UNIVERSITY OF CALIFORNIA SAN DIEGO

Algorithms for Statistical and Interactive Learning Tasks

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Computer Science

by

Christopher Tosh

Committee in charge:

Professor Sanjoy Dasgupta, Chair
Professor Ery Arias-Castro
Professor Russell Impagliazzo
Professor Lawrence Saul
Professor Jason Schweinsberg

2018

The Dissertation of Christopher Tosh is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

EPIGRAPH

We can only see a short distance ahead,
but we can see plenty there that needs to be done.

—Alan Turing

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am grateful to my advisor, Sanjoy Dasgupta, for his invaluable guidance throughout my time in graduate school. Working with Sanjoy has been a dream come true.

I would also like to thank all my friends at UCSD whose humor and insight has made these years so fun and interesting. Special mention goes to Akshay Balsubramani, Marco Carmosino, Suqi Liu, Stefanos Poulis, Sharad Vikram, and Michael Walter, who have all helped me in their own way.

Finally, I am forever indebted to my family and to my girlfriend, Allyson Cauble-Chantrenne. Their love and support has made me who I am today.

Chapter 2 contains material that has been submitted for publication as it may appear in Journal of Machine Learning Research. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

Chapter 3 contains material that is currently being prepared for submission for publication of the material. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

Chapter 5 contains material as it appears in "Lower bounds for the Gibbs sampler over mixtures of Gaussians." C. Tosh and S. Dasgupta. In International Conference of Machine Learning 2014. The dissertation author was the primary investigator.

Chapter 6 contains material as it appears in "Mixing Rates for the alternating Gibbs sampler over Restricted Boltzmann Machines and friends." C. Tosh. In International Conference of Machine Learning 2016. The dissertation author was the primary investigator.

Chapter 7 contains material as it appears in "Diameter-based active learning." C. Tosh and S. Dasgupta. In International Conference of Machine Learning 2017. The dissertation author was the primary investigator.

Chapter 8 contains material that is currently being prepared for submission for publication of the material. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

VITA

2012	Bachelor of Science, University of Texas at Austin
2018	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Algorithms for Statistical and Interactive Learning Tasks

by

Christopher Tosh

Doctor of Philosophy in Computer Science

University of California San Diego, 2018

Professor Sanjoy Dasgupta, Chair

In the first part of this thesis, we examine the computational complexity of three fundamental statistical tasks: maximum likelihood estimation, maximum a posteriori estimation, and approximate posterior sampling. We show that maximum likelihood estimation for mixtures of spherical Gaussians is NP-hard. We also demonstrate that, in many instances, hardness of maximum likelihood estimation implies hardness of maximum a posteriori estimation and approximate posterior sampling.

In the second part of this thesis, we explore the behavior of a common sampling algorithm known as the Gibbs sampler. We show that in the context of Bayesian Gaussian mixture models, this algorithm converges very slowly, even when the data looks as though it were generated by the model. We also demonstrate that when a particular variant of the Gibbs sampler is used in the context of a class of bipartite graphical models, called Restricted Boltzmann Machines, it can be guaranteed to converge quickly under certain conditions.

In the third part of this thesis, we consider learning problems in which the learner is allowed to solicit interaction from a user. In the context of classification, we present an efficient active learning algorithm whose performance is guaranteed to be competitive with that of any active learning algorithm for the particular instance under consideration. We also introduce a generic framework, termed interactive structure learning, for interactively learning complex structures over data, and we present a simple and effective algorithm that enjoys nice statistical properties in this setting.

Chapter 1

Introduction

Recent years have witnessed an explosion of models and applications emanating from the machine learning community. With each new model come new algorithmic questions: How hard is the model to fit in general? Under what assumptions do standard algorithms succeed? Can we utilize outside help to improve our fitting procedures? In this thesis, we explore each of these questions in depth.

1.1 The computational complexity of statistical tasks

When faced with a computational problem, one of the most basic algorithmic questions that arises is an existential one: 'do there exist efficient algorithms for this problem?' Positive answers are often quite straightforward and usually come in the form of an efficient algorithm. Negative answers, also known as hardness results, are often only conditionally negative, meaning the existence of efficient algorithms for this problem would imply some event that many computer scientists think unlikely.

The argument behind a negative answer generally comes in the form of an efficient algorithm, known as a reduction, that transforms an instance of a problem believed to be hard into an instance of the problem under consideration such that the resulting answer can be easily translated to an answer of the original problem. Perhaps the most famous group of problems that are used as the starting points of reductions is the class of NP-hard problems, due to the fact that they are widely believed to not admit efficient algorithms.

In this thesis, we will be mainly be concerned with reductions and hardness results for three computational tasks that arise in probabilistic modeling: maximum likelihood (ML) estimation of the model given data, maximum a posteriori (MAP) estimation when a prior distribution over models has been specified, and (approximate) sampling of the posterior distribution over models.

1.1.1 ML estimation

Consider, as a running example, the problem of fitting a mixture of spherical Gaussians to data. In this setting, we have a collection of data $x_1, \ldots, x_n \in \mathbb{R}^d$ that we model as having been generated from a mixture of k normal populations. That is, they are each independently and identically distributed (i.i.d.) according to

$$\pi_1 \, N(\mu_1, \sigma_1^2) + \cdots + \pi_k \, N(\mu_k, \sigma_k^2)$$

where $N(\mu, \sigma^2)$ is the multivariate spherical Gaussian distribution with density

$$N(x; \mu, \sigma^2) \ = \ \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}.$$

Given a data set, how do we go about recovering the relevant parameters, i.e. the μ's, π's, and σ's (which we will conveniently package up into a single vector as $(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma})$)?

One approach to this problem, known as the maximum likelihood approach, is to search for the parameters that make the data most probable. Formally, these are the parameters which maximize the likelihood, or equivalently the log-likelihood, of the data.

In the case of our mixture of Gaussians, the log-likelihood function is given by

$$LL(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}) \;=\; \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \, N(x_i; \mu_j, \sigma_j^2) \right).$$

The parameter set that maximizes the log-likelihood is called the maximum likelihood estimate, or ML estimate, and enjoys nice statistical properties [87, 57]. Though ML estimation assumes that the data is generated according to the probabilistic model in question, it is ultimately an optimization problem and, therefore, is well-defined for data sets that may not obey the distributional assumptions of the model. We will be interested in understanding the complexity of this optimization problem for specific probabilistic models, in particular for mixtures of spherical Gaussians and topic models.

1.1.2 MAP estimation and posterior sampling

The Bayesian approach to probabilistic modeling allows users to incorporate prior knowledge of the underlying parameters into the modeling process in the form of a distribution, called the *prior distribution* or sometimes simply the *prior*. That is, the unknown parameters are modeled as random variables whose a priori distribution is known. Bayes' rule tells us that we can simply multiply the prior distribution with the likelihood of the data to recover the posterior distribution of the parameters (up to a normalization constant). In the case of our mixtures of Gaussians model, if we say the prior distribution over our parameters follows a known density $q(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma})$, then the posterior distribution takes the following form.

$$\Pr(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma} \,|\, x_1, \ldots, x_n) \;\propto\; q(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}) \prod_{i=1}^{n} \sum_{j=1} \pi_j \, N(x_i; \mu_j, \sigma_j^2).$$

There are two tasks that immediately arise here: finding the parameters which maximize the posterior density, known as maximum a posteriori (or MAP) estimation, and sampling, perhaps only approximately, from the posterior distribution. The first of these, MAP estimation, may be seen as a natural Bayesian counterpart to ML estimation. The latter, posterior sampling, can be used to estimate various posterior probabilities and expectations of interest via Monte Carlo methods [42].

On the surface, both MAP estimation and posterior sampling seem at least somewhat related to ML estimation, since all three involve the likelihood of the data under the model. However, it is also true that without any restrictions on the prior, the posterior distribution may be concentrated on arbitrary points, meaning that the MAP estimate and samples from the posterior can differ significantly from the ML estimate. Such priors, however, are almost never used in practice.

In this thesis, we delve into the computational relationship of these problems. In particular, we explore cases where hardness of ML estimation for a particular probabilistic model implies hardness of the corresponding MAP estimation and approximate posterior sampling problems under commonly used priors.

1.2 Markov chains and mixing rates

As mentioned above, sampling from distributions is an important algorithmic primitive. It has found use not only in Bayesian statistics but also in theoretical computer science, statistical physics, and mathematical finance, among others. The widespread use of sampling algorithms is due, in part, to Monte Carlo methods, which utilize such samplers to approximate complicated integrals.

One classical way to sample from complicated distributions is via Markov chains. Formally, a Markov chain is a stochastic process $(X_t)_{t=0}^{\infty}$ that satisfies the *Markov property*:

$$\Pr(X_{t+1} \in A \mid X_t, \ldots, X_0) \;=\; \Pr(X_{t+1} \in A \mid X_t).$$

Informally, the Markov property asserts that given the value of X_t, the distribution of

X_{t+1} is independent of all previous values $X_0, \ldots, X_{t-1}$.

A fundamental result of Markov chain theory says that if a Markov chain satisfies certain properties, then it converges to a unique distribution, called its stationary distribution, regardless of its initial state [66, Theorem 4.9]. This powerful theorem tells us that if we can design a Markov chain with the appropriate stationary distribution, then it will eventually start producing samples from that distribution.

Convergence, however, is only guaranteed in the limit, meaning that for any particular finite time, our Markov chain may not actually be distributed according to the stationary distribution. Thus, we settle for approximation and instead hope that if we run our Markov chain for long enough, its distribution will be close to the stationary distribution. The rate at which this approximate convergence occurs is known as the *mixing rate*, and it is a central object of study in the theory of Markov chains.

In Part II, we investigate the mixing rates of commonly used Markov chains in particular cases of interest.

1.3 Interactive learning

In Part III, we shift our attention to settings in which a learner incorporates interaction into its fitting procedure. Such *interactive learning* settings come in a variety flavors that can roughly be characterized by the sources of interaction and whether or not the learner drives the interaction. In this thesis, we will focus on two settings in which a learner solicits relatively simple feedback from an intelligent user.

1.3.1 Active learning

In many situations where a classifier is to be learned, there is an abundance of unlabeled data, but labels are expensive to obtain. This occurs in wide variety of settings, including computer vision, bioinformatics, and natural language processing, where data collection and storage costs have decreased in recent years, but acquiring labels still requires humans to inspect the data points or experiments to be conducted. The pool-based active learning model is motivated by such scenarios.

In this model, a learning algorithm is presented with a large collection, or pool, of unlabeled examples, and adaptively queries certain data points for their labels. The hope is that by focusing on informative data points, an active learner can find a low error hypothesis with fewer labels than a passive learner that relies on labels provided at random. To see the potential savings of such a scheme, we turn to an example.

Example: Thresholds

Consider the task of learning a one-dimensional threshold. In this setting, our data points $x_1, \ldots, x_n$ are i.i.d. draws from some distribution $\mathcal{D}$ over the real line $\mathbb{R}$ and their labels $y_1, \ldots, y_n$ take values in $\{-1, +1\}$. There is some threshold $t^* \in \mathbb{R}$ that induces these labels, i.e.

$$y_i = \begin{cases} +1 & \text{if } x_i \geq t^* \\ -1 & \text{if } x_i < t^* \end{cases} \tag{1.1}$$

Our goal in this setting is to find a threshold $t \in \mathbb{R}$ that will have low error relative to t^*:

$$\operatorname{err}(t) \;=\; \Pr_{x \sim \mathcal{D}}(\mathbf{1}[x \geq t^*] \neq \mathbf{1}[x \geq t])$$

where $\mathbf{1}[\cdot]$ is the indicator function that is 1 if the condition holds true and 0 otherwise.

If we are presented with all the labels $y_1, \ldots, y_n$, then we know that with high probability, any threshold consistent with these labels will have error bounded by $O(1/n)$. Something stronger actually holds here: to find a threshold with error ϵ, it is necessary and sufficient for a passive learner to receive $O(1/\epsilon)$ labels.

It turns out that an active learning algorithm can get away with fewer labels in this setting. To see why, note that if we have the label of y_i, equation (1.1) can be reversed to tell us something about t^*:

$$\begin{aligned} t^* &\leq x_i \quad \text{if} \quad y_i = +1 \\ t^* &> x_i \quad \text{if} \quad y_i = -1 \end{aligned}$$

In particular, if y_i is +1, we know that $y_j = +1$ for all $x_j \geq x_i$; and if y_i is -1, we know that $y_j = -1$ for all $x_j \leq x_i$. An active learner can use binary search to exploit this structure and recover all n labels after $O(\log n)$ queries. Thus, with $O(1/\epsilon)$ unlabeled data points and $O(\log 1/\epsilon)$ labels, an active learner can find a classifier with error ϵ. With respect to the number of labels required, this is an exponential improvement over the passive learner!

As encouraging as our threshold example is, we cannot hope to always do exponentially better than the passive learner, even in the noiseless and realizable setting [31]. A more realistic goal is to construct an active learner whose performance is comparable to the best active learning algorithm for the scenario under consideration. We will see that under certain assumptions this is not only theoretically possible, but there is a simple and efficient algorithm that achieves this.

1.3.2 Interactive structure learning

Consider the problem of fitting a structure, such as a flat clustering or a hierarchy, to some particular data set. There are many ways to do this, but they typically involve two discrete steps. The first step is to construct a cost function, define a probabilistic model, or derive a posterior distribution over structures, and the second is to optimize the cost function, fit the model, or sample from the posterior.

There have been many advances in recent years for doing these tasks with no supervision, but it is rarely the case that the structures produced by such procedures perfectly align with what a user expects or has in mind. There are a many reasons why

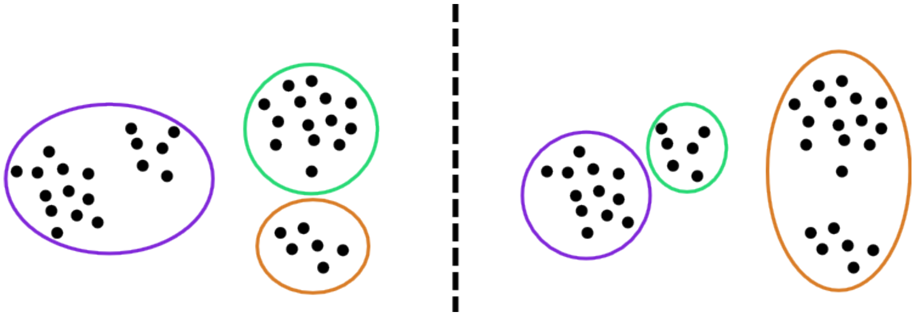


Figure 1.1. An ambiguous data set with two valid clusterings.

such deficiencies arise. Here we highlight three important ones.

- *Computational reasons.* It may be that optimizing the cost function or sampling from the posterior is difficult in general, and the only available algorithms find crude approximations of the optimum.

- *Modeling issues.* The data may not exactly match the assumptions of the model a user has in mind. For example, a user might try to fit a mixture of Gaussians to data that was generated by a mixture of distributions with heavier tails.

- *Ambiguous data.* Complex data can be organized in a multitude of reasonable ways, particularly in high-dimensions, but a user may only really be satisfied with some small subset of these. Consider the two clusterings of the same toy data set in Figure 1.1. Both could conceivably be considered legitimate ways to cluster the data set, but if a user only views one as correct, how is an unsupervised method to choose?

Interactive structure learning addresses such situations by allowing users to iteratively provide feedback on small subsets of a potential structure. As in the active learning scenario, learners in this setting pose queries to a user. But interactive structure learning differs from, or rather generalizes, active learning in two key ways.

- *Mini-structures as queries.* In the traditional active learning setting, a learner asks questions of the form 'what is the label of this specific point?' In interactive structure learning, questions are of the form ‘does this structure on these k points look correct?’ In soliciting feedback on several points at once, this type of query provides context to a user, and may result in higher quality feedback.

- *Partial correction feedback.* In active learning, a user provides a single label for the point that was queried. In interactive structure learning, a user corrects or confirms some aspect of the structure that was presented. For example, if a user is presented with a clustering of ten points, they may respond with ‘no, those two points should not be in the same cluster’ or ‘yes, those four points definitely should be clustered together.’

These changes allow for natural feedback in a wide variety of structure learning tasks, but they also present challenges not present in the traditional active learning setting: When queries are no longer single points but rather entire structures, how do we measure the information content of a query? In allowing users to only provide partial feedback on structures, how can we guarantee consistency? Can we construct generic algorithms that operate in any interactive structure learning environment? In this thesis, we tackle each of these questions.

1.4 Summary of results

In Part I, we will be concerned with the computational complexity of various statistical tasks. In particular, Chapter 2 demonstrates that estimating the maximum likelihood solution for a mixture of spherical Gaussians is NP-hard, and Chapter 3 gives generic reductions from maximum likelihood estimation to two canonical Bayesian tasks: maximum a posteriori estimation and approximate posterior sampling. Taken together, these two chapters show that approximate sampling from the posterior of a Bayesian mixture of Gaussians model is NP-hard in general.

In Part II, we will investigate the mixing rates of a specific Markov chain, known as the Gibbs sampler, in two distinct contexts. In Chapter 5, we look at the Gibbs sampler in the Bayesian mixture of Gaussians setting and show that its mixing rate can be quite slow, even when the data is well-modeled by a mixture of Gaussians. In Chapter 6, we demonstrate that when a particular variant of the Gibbs sampler is used in the context of a class of bipartite graphical models, called Restricted Boltzmann Machines (RBMs), its mixing rate can be guaranteed to be fast in certain instances. Chapter 6 also provides lower bounds for the same chain in other instances.

In Part III, we present some interactive learning algorithms. In Chapter 7, we give an efficient active learning algorithm whose label complexity is close to the best achievable by any active learning algorithm in the noiseless and realizable setting. In Chapter 8, we present the interactive structure learning framework, and provide a simple, efficient, and generic algorithm. We show that it enjoys nice theoretical guarantees, even in the presence of noise, and performs well empirically.

Part I

Computational complexity of statistical tasks

Chapter 2

Maximum likelihood estimation is NP-hard for mixtures of spherical Gaussians

In this chapter, we investigate the computational complexity of fitting a particular probabilistic model, mixtures of spherical Gaussians. In particular, we will show that finding the parameters which approximately maximize the log-likelihood is an NP-hard problem.

2.1 Mixtures of Gaussians

A *spherical Gaussian* in $\mathbb{R}^d$ is a distribution specified by its mean $\mu \in \mathbb{R}^d$ and variance $\sigma^2 > 0$, with density

$$N(x; \mu, \sigma^2) \; = \; \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right).$$

(The standard notation for this Gaussian is $N(\mu, \sigma^2 I_d)$, but we will drop the identity matrix as a shorthand.)

When data arise from several sources, or form several clusters, it is common to model each source or cluster by a spherical Gaussian. If there are k sources, the resulting overall distribution is a mixture of k Gaussians,

$$\pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2) + \cdots + \pi_k N(\mu_k, \sigma_k^2),$$

where $\mu_i \in \mathbb{R}^d$ and σ_i^2 are the mean and variance of the ith component, and π_i is the fraction of the distribution that arises from this component. In what follows, we will often package the parameters together as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$.

A standard statistical task is to fit a mixture of k Gaussians to a given data set. There are many approaches to doing so, but a common formulation is as an optimization problem [38], where given data points $x_1, \ldots, x_n \in \mathbb{R}^d$, the goal is to find the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ that maximize the *log-likelihood*

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma_j^2) \right). \tag{2.1}$$

In this chapter, we establish the computational hardness of this estimation problem. This is in contrast to recent positive results [17, 72, 56, 52] that provide efficient algorithms when the input data is in fact generated from a Gaussian mixture.

2.1.1 Gaussians with the same variance

We start with the simplest subcase, where the variances of the components are constrained to be the same.

> MIXTURES OF SPHERICAL GAUSSIANS WITH SAME VARIANCE: MOG-SV
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k; unary parameter b.
> *Output*: A mixture of k spherical Gaussians with the same variance, $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, whose log-likelihood
>
> $$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma^2) \right).$$
>
> is within an additive factor $1/b$ of optimal.

The input parameter b specifies the desired precision of the solution.

MOG-SV is similar to the k-means clustering problem, which is NP-hard [2].

k-MEANS
Input: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k.
Output: A collection of k "centers" $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ in $\mathbb{R}^d$ that minimize the cost function

$$\Phi(\boldsymbol{\mu}) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - \mu_j\|^2.$$

The biggest difference between the two problems is that k-means assigns each data point x_i to a single center μ_j (a "hard" clustering), while the mixture of Gaussians effectively spreads it out over all the centers (a "soft" clustering). Earlier work [5] has established that a "hard clustering" version of the mixture of Gaussians problem is NP-hard. Here we consider the more standard formulation, and show that it is hard even when $k = 2$.

Theorem 2.1. MOG-SV *is NP-hard on instances with* $k = 2$.

The proof follows from the observation that an additive approximation to the best MOG-SV solution yields a multiplicative approximation to the best k-means solution:

Lemma 2.2. *Fix any data set* $x_1, \ldots, x_n \in \mathbb{R}^d$ *and any positive integer* k. *Let* LL_{OPT} *denote the log-likelihood of the optimal solution to* MOG-SV, *and* Φ_{OPT} *the lowest achievable* k*-means cost. For any parameters* $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, *we have*

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \;\leq\; \frac{4 \ln k}{d} + \frac{2}{nd} \left(LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \right).$$

The first term on the right-hand side comes from the discrepancy between hard and soft clustering. It can be made negligible by increasing the dimension, for instance by padding each point with extra zero-valued coordinates.

Lemma 2.2 can also be combined with a recent hardness of approximation result for k-means [9] to show that, if k is allowed to be large, MOG-SV cannot be approximated within an additive factor of $o(nd)$.

Theorem 2.3. *There is a family of* MOG-SV *instances with the following properties:*

- *An instance with n points has dimension $O(n)$.*
- *Each point is $\{0,1\}$-valued and has $O(1)$ nonzero coordinates.*
- $k = \Theta(n)$.

For some absolute constant c_o, it is NP-hard to approximate MOG-SV *on such instances within an additive factor of $c_o dn$.*

The specific form of this result (additive versus multiplicative approximation, importance of interpoint distances) is motivated by the unusual properties of the log-likelihood objective. To begin with, consider the problem of fitting a single Gaussian to a data set $\mathcal{X} \subset \mathbb{R}^d$ of size n. A quick calculation shows that the log-likelihood (of the maximum likelihood estimate) is

$$\frac{dn}{2} \ln \frac{d}{2\pi e} - \frac{dn}{2} \ln \operatorname{radius}(\mathcal{X}), \quad \textit{where} \;\; \operatorname{radius}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \|x - \operatorname{mean}(\mathcal{X})\|^2.$$

Depending on the scale of the data, this log-likelihood could be positive, negative, or zero. When fitting a mixture of k Gaussians, the log-likelihood has a term of this sort for each cluster, plus an additional term of size $\pm n \ln k$ due to the mixing weights. For the kind of instance described in the theorem, any cluster with at least two points has radius $\Theta(1)$ and thus the log-likelihoods of all reasonable mixture models lie in an interval of size $O(dn)$.

The proofs of these results appear in Section 2.2.

2.1.2 Gaussians with differing variances

When the different Gaussian components are allowed to have different variances, and $k > 1$, the maximum-likelihood solution is always degenerate. This is because it is possible to make the log-likelihood go to infinity by centering one of the Gaussians at a single data point and letting its variance go to zero. Thus, in order for the problem to be well-defined, an additional constraint must be introduced. One option is to force all variances to be non-negligible.

> MIXTURES OF SPHERICAL GAUSSIANS WITH CONSTRAINED VARIANCES: MOG
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k; value $\sigma_o > 0$; unary integer b.
> *Output*: A mixture of k spherical Gaussians $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is within an additive factor $1/b$ of optimal, subject to the constraint $\sigma_1, \ldots, \sigma_k \geq \sigma_o$.

This problem is slightly further from k-means, but remains intractable.

Theorem 2.4. MOG *is NP-hard on instances with* $k = 2$.

The proof appears in Section 2.3.

2.2 Mixtures of spherical Gaussians with the same variance

2.2.1 Induced partitions

We start with a basic relation between hard and soft clustering that applies to arbitrary mixture models, not just those with Gaussian components of the same variance.

Although a mixture model represents a soft clustering, it also induces a natural hard partition. For data set $\mathcal{X}$ and mixture of Gaussians $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, this hard partition has clusters

$$\mathcal{X}_j \;=\; \left\{ x \in \mathcal{X} \,:\, j = \underset{\ell}{\operatorname{argmax}} \;\pi_\ell N(x; \mu_\ell, \sigma_\ell^2) \right\} \tag{2.2}$$

(breaking ties arbitrarily). The log-likelihood of a mixture is easily bounded in terms of the likelihood of the corresponding hard partition.

Lemma 2.5. *Pick any mixture* $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ *and data set* $\mathcal{X} = \{x_1, \ldots, x_n\}$.

(a) For any partition $(\mathcal{X}'_1, \ldots, \mathcal{X}'_k)$ of $\mathcal{X}$, we have

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;\geq\; \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \ln(\pi_j N(x; \mu_j, \sigma_j^2)).$$

(b) For the partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ induced by $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, as in Eq (2.2), we have

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;\leq\; n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \ln(\pi_j N(x; \mu_j, \sigma_j^2)).$$

Proof. Recall from (2.1) that the contribution of each data point x_i to $LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is

$$\ln\left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma_j^2)\right).$$

For $x_i \in \mathcal{X}'_j$, we can lower-bound this contribution by $\ln(\pi_j N(x_i; \mu_j, \sigma_j^2))$. Similarly, if $x_i \in \mathcal{X}_j$, then we can upper-bound the contribution by $\ln(k\pi_j N(x_i; \mu_j, \sigma_j^2))$, by the manner in which the hard partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ is defined. □

2.2.2 Proof of Lemma 2.2

As in the statement of Lemma 2.2, fix data $x_1, \ldots, x_n \in \mathbb{R}^d$, and define LL_{OPT} to be the log-likelihood of the optimal solution of MOG-SV. Let Φ_{OPT} be the optimal k-means cost.

Pick any parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, and let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the induced hard partition of the data set, as per Eq (2.2). From Lemma 2.5,

$$\begin{aligned} LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) &\leq n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln \pi_j + \frac{d}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\|x - \mu_j\|^2}{2\sigma^2}\right) \\ &\leq n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \end{aligned}$$

$$
\begin{aligned}
&\leq n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\Phi(\boldsymbol{\mu})}{2\sigma^2} \\
&\leq n \ln k + \frac{nd}{2} \ln\left(\frac{nd}{2\pi\Phi(\boldsymbol{\mu})}\right) - \frac{nd}{2},
\end{aligned}
$$

where the last inequality comes from solving for the optimal value of σ^2 (namely, $\Phi(\boldsymbol{\mu})/nd$) in the preceding line.

Suppose the optimal k-means solution is realized by centers $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_k^*)$. Let $\pi_1^* = \cdots = \pi_k^* = 1/k$ and $\sigma^{*2} = \Phi(\boldsymbol{\mu}^*)/nd$. To bound the log-likelihood of the mixture model $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*)$, we look at the hard partition that it induces, $(\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*)$, and notice that $\mathcal{X}_j^*$ consists of points whose closest center is μ_j^*. We then apply Lemma 2.5 to get

$$
\begin{aligned}
LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*) &\geq \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \left(\ln \pi_j^* + \frac{d}{2} \ln\left(\frac{1}{2\pi\sigma^{*2}}\right) - \frac{\|x - \mu_j^*\|^2}{2\sigma^{*2}} \right) \\
&= -n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^{*2}}\right) - \frac{1}{2\sigma^{*2}} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \|x - \mu_j^*\|^2 \\
&= -n \ln k + \frac{nd}{2} \ln\left(\frac{1}{2\pi\sigma^{*2}}\right) - \frac{1}{2\sigma^{*2}} \Phi(\boldsymbol{\mu}^*) \\
&= -n \ln k + \frac{nd}{2} \ln\left(\frac{nd}{2\pi\Phi(\boldsymbol{\mu}^*)}\right) - \frac{nd}{2},
\end{aligned}
$$

where the last equality comes from substituting in the value of σ^{*2}. Combining our bounds for the two mixtures, we get

$$
\begin{aligned}
LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) &\geq LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \\
&\geq \frac{nd}{2} \ln\left(\frac{\Phi(\boldsymbol{\mu})}{\Phi(\boldsymbol{\mu}^*)}\right) - 2n \ln k.
\end{aligned}
$$

Rearranging terms yields the lemma statement.

2.2.3 Proof of Theorem 2.1

With Lemma 2.2 in place, a reduction from k-means to MOG-SV is almost immediate. There are various hardness results available for k-means [2, 32, 71, 9]; of these, we use [2] as a starting point.

Theorem 2.6 ([2])**.** *There exists a family of k-means instances with the following properties, for some low-order polynomials $\alpha(\cdot)$ and $\beta(\cdot)$:*

- *For an instance containing n points, each point has dimension at most $\alpha(n)$, with individual coordinates taking values in $\{-1, 0, 1\}$.*
- *It is NP-hard to approximate the best k-means solution, with $k = 2$, within a factor of $1 + 1/\beta(n)$.*

To prove Theorem 2.1, we reduce the problem of finding a $(1+1/\beta(n))$-approximate k-means solution to MOG-SV. Given an instance $x_1, \ldots, x_n$ of k-means:

- Pad each point with additional zero-valued coordinates until the dimension d exceeds $16\beta(n) \ln k$. This has no effect on interpoint distances or on the optimal k-means cost.
- Solve MOG-SV for these modified points, with precision parameter $b = 1$. This yields $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ such that $LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \leq 1$, where LL_{OPT} is the optimal log-likelihood. It follows from Lemma 2.2 that

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \leq \frac{4 \ln k}{d} + \frac{2}{nd} \leq \frac{1}{2\beta(n)},$$

  whereupon $\Phi(\boldsymbol{\mu}) \leq \Phi_{OPT}(1 + 1/\beta(n))$.

2.2.4 Proof of Theorem 2.3

A recent hardness of approximation result for k-means shows the following.

Theorem 2.7 ([9]). *There is a family of k-means instances with the following properties:*

- *An instance with n points has dimension at most n, points that are $\{0,1\}$-valued (and have at most two non-zero coordinates), and a target number of clusters $k = \Omega(n)$.*
- *It is NP-hard to approximate the optimal k-means solution within a factor c, for some absolute constant $c > 1$.*

Pick any $c_o < (1/2)\ln c$. To see that it is hard to approximate MOG-SV within an additive factor $c_o nd$, we reduce from k-means as follows. Start with an instance $x_1, \ldots, x_n \in \mathbb{R}^d$ of the type described in Theorem 2.7. Then:

- If necessary, pad points with zero-valued coordinates to bring the dimension up to

$$d \geq \frac{4 \ln k}{(\ln c) - 2c_o}.$$

- Obtain an approximate solution $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ to MOG-SV on these points such that $LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \leq c_o nd$.
- Return the centers $\boldsymbol{\mu}$.

By Lemma 2.2, we have

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \;\leq\; \frac{4 \ln k}{d} + \frac{2}{nd} c_o nd \;\leq\; \ln c,$$

so that $\boldsymbol{\mu}$ is a c-approximate solution to the k-means instance.

2.3 The general case

We now consider the case where the variances are allowed to differ but are uniformly lower bounded. Specifically, a mixture model $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is *admissible* if all $\sigma_j \geq \sigma_o$, where σ_o is supplied as part of the input.

The basic reduction still applies, with an additional device to force all variances to be close to the lower bound—and therefore approximately equal.

2.3.1 Controlling the variances

Lemma 2.8. *Fix any data set $\mathcal{X} = \{x_1, \ldots x_n\}$ in $\mathbb{R}^d$, and let $D = \max_{i \neq i'} \|x_i - x_{i'}\|$ denote its diameter. Pick any $\Delta, \delta > 0$. If the dimension d satisfies*

$$d \geq \frac{4}{\delta}\left(\frac{nD^2}{2\sigma_o^2} + n \ln k + \Delta\right), \tag{2.3}$$

then any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ within an additive factor Δ of optimal (that is, $LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \geq LL_{OPT} - \Delta$) has the following property: in the associated hard partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$, any nonempty cluster $\mathcal{X}_j$ has $\sigma_j^2 \leq \sigma_o^2(1+\delta)$.

Proof. Pick any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ that is within Δ of optimal, and let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the associated hard partition. Let $\widetilde{\mu}_j$ denote the cluster means:

$$\widetilde{\mu}_j \; = \; \frac{1}{|\mathcal{X}_j|} \sum_{x \in \mathcal{X}_j} x.$$

Using Lemma 2.5, we can compare the log-likelihood of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ to that of the adjusted parameters $(\boldsymbol{\pi}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}})$, where each $\widetilde{\sigma}_j = \sigma_o$.

$$\begin{aligned}
&LL(\boldsymbol{\pi}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}}) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&\geq \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln(\pi_j N(x; \widetilde{\mu}_j, \sigma_o^2)) - \ln(\pi_j N(x; \mu_j, \sigma_j^2))\right) - n \ln k \\
&= \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\frac{d}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{\|x - \widetilde{\mu}_j\|^2}{2\sigma_o^2} - \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^2} + \frac{\|x - \mu_j\|^2}{2\sigma_j^2}\right) - n \ln k \\
&= \sum_{j=1}^{k} \left(\frac{d|\mathcal{X}_j|}{2} \ln \frac{\sigma_j^2}{\sigma_o^2} + \sum_{x \in \mathcal{X}_j} \left(\frac{\|x - \mu_j\|^2}{2\sigma_j^2} - \frac{\|x - \widetilde{\mu}_j\|^2}{2\sigma_o^2}\right)\right) - n \ln k
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{j=1}^{k} \left(\frac{d|\mathcal{X}_j|}{2} \ln \frac{\sigma_j^2}{\sigma_o^2} + \left(\frac{1}{2\sigma_j^2} - \frac{1}{2\sigma_o^2} \right) \sum_{x \in \mathcal{X}_j} \|x - \widetilde{\mu}_j\|^2 \right) - n \ln k \\
&\geq \sum_{j=1}^{k} |\mathcal{X}_j| \left(d \ln \frac{\sigma_j}{\sigma_o} - \frac{D^2}{2\sigma_o^2} \right) - n \ln k \quad \geq \quad d \ln \frac{\max_{j:\mathcal{X}_j \neq \emptyset} \sigma_j}{\sigma_o} - \frac{nD^2}{2\sigma_o^2} - n \ln k.
\end{aligned}
$$

In the second-last line, we have exploited the fact that $\widetilde{\mu}_j$ is the mean of cluster $\mathcal{X}_j$, so that $\sum_{x \in \mathcal{X}_j} \|x - \widetilde{\mu}_j\|^2 \leq \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2$, and for the last line we have used $\|x - \widetilde{\mu}_j\| \leq D$.

The difference above is at most Δ, and thus for each nonempty cluster $\mathcal{X}_j$,

$$
d \ln \frac{\sigma_j}{\sigma_o} - \frac{nD^2}{2\sigma_o^2} - n \ln k \leq \Delta,
$$

whereupon $\sigma_j^2 \leq \sigma_o^2(1 + \delta)$ given the bound (2.3) on the dimension d. □

This observation allows us to prove following analog of Lemma 2.2.

Lemma 2.9. *Following the terminology of Lemma 2.8, pick $\delta, \Delta > 0$ and suppose that the dimension satisfies (2.3). Pick any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is within an additive factor Δ of the optimal. Then*

$$
\Phi(\boldsymbol{\mu}) \;\leq\; (1 + \delta) \left(2\sigma_o^2(\Delta + 2n \ln k) + \Phi_{OPT} \right).
$$

Proof. Let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the hard partition of the data set induced by $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. By Lemma 2.8, we know that for any nonempty cluster $\mathcal{X}_j$, the variance σ_j^2 is at most $(1+\delta)\sigma_o^2$. Thus, using Lemma 2.5, we have

$$
\begin{aligned}
LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &\leq n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln \pi_j + \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^2} - \frac{\|x - \mu_j\|^2}{2\sigma_j^2} \right) \\
&\leq n \ln k + \sum_{j=1}^{k} \left(\frac{|\mathcal{X}_j| d}{2} \ln \frac{1}{2\pi\sigma_j^2} - \frac{1}{2\sigma_j^2} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \right)
\end{aligned}
$$

$$\le n \ln k + \frac{nd}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{1}{2(1+\delta)\sigma_0^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2$$
$$\le n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2}\right) - \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_o^2}$$

Let $\mu_1^*, \ldots, \mu_k^*$ be an optimal k-means solution and let $(\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*)$ be the hard partition of the data set induced by the mixture model $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$ where $\pi_j^* = 1/k$ and $\sigma_j^* = \sigma_o$ for all j. Again using Lemma 2.5,

$$\begin{aligned} LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*) &\ge \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \left(\ln \pi_j^* + \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^{*2}} - \frac{\|x - \mu_j^*\|^2}{2\sigma_j^{*2}} \right) \\ &= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2}\right) - \frac{\Phi(\boldsymbol{\mu}^*)}{2\sigma_o^2} \end{aligned}$$

Then by the near-optimality of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, we have

$$\begin{aligned} \Delta &\ge LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &\ge LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &\ge \left(-n \ln k + \frac{nd}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{\Phi(\boldsymbol{\mu}^*)}{2\sigma_o^2} \right) - \left(n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2}\right) - \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_0^2} \right) \\ &= \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_o^2} - \frac{\Phi_{OPT}}{2\sigma_o^2} - 2n \ln k \end{aligned}$$

Rearranging gives the theorem statement. □

2.3.2 Proof of Theorem 2.4

Once again we reduce from k-means, using the hardness result of [2], summarized in Theorem 2.6. Recall that the family of instances for which k-means was shown to be hard has $k = 2$, $d = \text{poly}(n)$, and points with $\{-1, 0, 1\}$-valued coordinates.

Starting with such an instance $x_1, \ldots, x_n \in \mathbb{R}^d$, we show how MOG can be used to find an $(1 + 1/\beta(n))$-approximate solution to k-means.

- Let D denote the diameter of the points; it is polynomial in n.
- Set $\delta = 1/(5\beta(n))$ and
$$\sigma_o^2 = \frac{\delta}{2(1+2n\ln k)}.$$
- Pad the points with zero-valued coordinates to bring the dimension up to at least
$$d = \frac{4}{\delta}\left(\frac{nD^2}{2\sigma_o^2} + n\ln k + 1\right).$$
- Invoke MOG on these modified points, with target precision $b = 1$ and variance lower bound σ_o^2. This returns a mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is at least $LL_{OPT} - 1$, subject to the variance constraint.
- Return centers $\boldsymbol{\mu}$.

Lemma 2.9, with $\Delta = 1$, asserts that

$$\Phi(\boldsymbol{\mu}) \;\leq\; (1+\delta)(2\sigma_o^2(1+2n\ln k) + \Phi_{OPT}) \;\leq\; (1+\delta)(\delta + \Phi_{OPT}) \;\leq\; (1+5\delta)\Phi_{OPT},$$

which is at most $(1 + 1/\beta(n))\Phi_{OPT}$. For the last inequality, we have used the fact that $\Phi_{OPT} \geq 1/2$ since all interpoint distances are ≥ 1.

Chapter 2 contains material that has been submitted for publication as it may appear in Journal of Machine Learning Research. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

Chapter 3

The relative complexity of maximum likelihood estimation, MAP estimation, and approximate posterior sampling

We saw in the previous chapter that finding the maximum likelihood estimate of a particular probabilistic model is an NP-hard problem. In this chapter, we look at what such types of results imply in a Bayesian context. In particular, we will show that when our models are well-behaved, certain Bayesian computational tasks are no easier computationally than maximum likelihood estimation.

3.1 Canonical learning tasks

When learning a probabilistic model, there are three computational tasks that commonly arise: maximum likelihood (ML) estimation of the model given data, maximum a posteriori (MAP) estimation when a prior distribution over models has been specified, and (approximate) sampling of the posterior distribution over models. We are interested in the relative *computational* complexity of these three tasks: what does the hardness—or tractability—of one imply about the others?

At a high level, MAP estimation is rather like ML estimation, with the added complication of the prior—and the two are known to converge to the same limit with infinite data, under certain conditions. Thus one would intuitively expect the MAP problem to be at least as hard as the ML problem.

The situation of approximately sampling is not as immediately clear. Sampling is known to be as hard as optimization for various statistical physics models with a "temperature" parameter: this temperature can be adjusted so that a sampler is essentially forced to return optimal or near-optimal solutions. In the usual setting of probabilistic learning, however, there is no such temperature knob. Nonetheless, there are other ways to produce a similar effect, and thus one would again expect, intuitively, that approximate sampling is no easier than ML estimation.

In this work, we make these intuitions precise. Considering probabilistic models in broad generality, we give simple conditions under which approximate MAP estimation and approximate posterior sampling can be shown to be at least as hard as approximate ML estimation. A key challenge here is formalizing issues of numerical precision.

We then illustrate these general reductions in two cases of interest. Starting from hardness results for maximum-likelihood estimation of Gaussian mixture models, which we demonstrated in Chapter 2, and topic models [4], we show how in both settings, the hardness extends also to MAP estimation and approximate sampling.

3.1.1 Numerical precision

When discussing standard combinatorial optimization problems such as set cover or maximum cut, the first step is to consider the *exact* version of the problem and, when that proves intractable, to consider *approximate* solutions. For such problems, the exact solution lies in a discrete space and is of polynomial size, but is difficult to find.

By contrast, many of the problems that we have in mind—such as estimation of Gaussian mixture models or of topic distributions—take solutions in continuous spaces. The exact optimal solutions may therefore not have compact representations. The following lemma illustrates how this can happen even for extremely simple models.

Lemma 3.1. *When fitting a mixture model $\frac{1}{2}N(-\mu, 1) + \frac{1}{2}N(\mu, 1)$ to the data set of three points $\{-2, 0, 2\}$, the maximum-likelihood choice of μ is irrational.*

Proof. Writing out the log-likelihood function,

$$\begin{aligned}\ln p(-2, 0, 2 \mid \mu) &= \ln\left(\frac{1}{2\sqrt{2\pi}}e^{-\mu^2/2} + \frac{1}{2\sqrt{2\pi}}e^{-\mu^2/2}\right) \\ &\quad + 2\ln\left(\frac{1}{2\sqrt{2\pi}}e^{-(2-\mu)^2/2} + \frac{1}{2\sqrt{2\pi}}e^{-(2+\mu)^2/2}\right) \\ &= -2\ln(2\sqrt{2\pi}) - 4 - \frac{3\mu^2}{2} + 2\ln\left(e^{2\mu} + e^{-2\mu}\right).\end{aligned}$$

Taking derivatives of the log-likelihood equation with respect to μ,

$$\frac{d}{d\mu}\ln p(-2, 0, 2 \mid \mu) = -3\mu + 4\tanh(2\mu).$$

This has two non-negative roots, one of which is zero. Evaluating the second derivative at zero, we have

$$\frac{d^2}{d^2\mu}\ln p(-2, 0, 2 \mid \mu)|_{\mu=0} = 8\mathrm{sech}^2(0) - 3 = 5.$$

Thus zero is a local minimum. Because $-3\mu + 4\tanh(2\mu)$ tends to $-\infty$ as μ goes to ∞, we can conclude that the other nonnegative root is the maximum likelihood estimate. Expanding the tanh, we see that this root satisfies $(3\mu - 4)e^{2\mu} + (3\mu + 4)e^{-2\mu} = 0$. By the Lindemann-Weierstrass theorem [12, chapter 1], the ML estimate must be irrational. □

Thus, exact solutions are ruled out from the very beginning, and we are forced to restrict ourselves to approximate versions of ML and MAP estimation. In particular, we need to characterize the quality of polynomial-sized solutions.

The case of sampling is even more challenging, since we cannot hope to sample exactly from continuous domains. The usual distance metric between the target distribution μ over a space Θ and the distribution ν from which samples are actually drawn is *total* *variation distance*:

$$d_{TV}(\mu, \nu) \;=\; \sup_{\text{measurable } A \subset \Theta} |\mu(A) - \nu(A)|.$$

For instance, this is the standard metric for assessing the convergence rates of Markov chains. In cases where μ has continuous support, ν must still be discrete because samples must be bounded in size, and thus this distance will be identically 1.

To overcome this, we introduce a generalization of total variation distance that takes the supremum over a subfamily of measurable sets that captures Θ *at a certain granularity*. We show how to construct suitable such families from ϵ-covers of Θ; our construction may be useful in other contexts.

3.1.2 Overview of results

In Section 3.3, we show that under conditions on the prior, there is a generic polynomial-time reduction from ML estimation to MAP estimation.

In Section 3.4, we define a notion of approximate posterior sampling that makes sense in continuous domains and we then give a generic reduction from ML estimation to this problem, again under conditions on the prior.

Sections 3.3 and 3.4 extend the hardness of many important ML estimation problems to their Bayesian counterparts provided that the prior meets certain mild conditions. In these cases, we cannot hope for efficient MAP estimation algorithms or rapidly mixing Markov chains unless the data is specially constrained or $NP = RP$.

Throughout our exposition, topic modeling serves as a running example. In particular, we extend a hardness result [4] for ML estimation of topic models to the corresponding Bayesian estimation problems. In Section 3.5, we do this also for the problem of estimating mixtures of Gaussians.

3.1.3 Methodology

Our goal is to reduce arbitrary maximum-likelihood (ML) estimation problems to their Bayesian counterparts. Without any particular knowledge about the specific model under consideration, we opt for the simplest type of reduction: duplication.

Let $\mathcal{P} = \{p(\cdot \,|\, \theta) \,:\, \theta \in \Theta\}$ be a family of parameterized probability densities and let a q_0 be a prior density over Θ. Consider a data sequence $X = (x_1, \ldots, x_n)$ and suppose that we replicate this sequence k times, i.e. we make k copies $X^{(i)} = (x_1, \ldots, x_n)$. Now what does the posterior distribution look like, given $X^{(1)}, \ldots, X^{(k)}$? It is of the form

$$\frac{1}{Z} q(\theta) p(X^{(1)}, \ldots, X^{(k)} \,|\, \theta) = \frac{1}{Z} q(\theta) \prod_{i=1}^{k} p(X^{(i)} \,|\, \theta) = \frac{1}{Z} q(\theta) \left(p(x_1, \ldots, x_n \,|\, \theta)\right)^k .$$

Here Z is the normalizing constant to make the density integrate to one. By simply replicating the data, we get an exponential increase on the weight of the likelihood function over the prior distribution. Thus, our general strategy will be to replicate the data until the posterior distribution is suitably concentrated around the maximum likelihood estimate.

The success of this approach ultimately hinges on the relationship of the prior distribution and the maximum likelihood estimate. If the prior distribution has very low density, say double-exponentially small, on parameters that are close to the maximum likelihood estimate, then to get large enough posterior weight on these parameters requires us to duplicate the data a very large number of times, certainly more than polynomial. On the other hand, many prior distributions, especially those over unbounded parameter spaces, do put very small weight on *some* parameters. Thus our data duplication technique ought to fail for instances where the maximum likelihood estimate lies in some small prior density region. How do we get around this?

The key observation is that many hardness reductions to ML estimations problems do not produce arbitrary instances. Indeed, these reductions often create instances with a large degree of regularity. And in many of these cases, the maximum likelihood solutions to these highly regular instances are themselves well-structured and, as a consequence, often have non-negligible weight, or are near other solutions with non-negligible weight, under many commonly-considered prior distributions. The upshot, then, is that if we restrict ourselves to reducing from instances that are known to be hard, then in many cases we can avoid the only obstacle to our duplication technique.

3.1.4 Related work, including connection with statistical literature

Of the computational tasks discussed in this work, ML estimation has seen the lion's share of hardness results [24, 49]. This is possibly because ML estimation does not have the additional complication of a prior distribution and can be easier to work with than MAP estimation and sampling.

In the computational literature, several algorithmic connections have been made between sampling and optimization. In [63], it was shown that simulated annealing, a technique that involves approximately sampling from a sequence of distributions, can be used for certain convex optimization problems. In [21], Langevin dynamics, a technique in which Gaussian noise is added to each step of gradient descent, was used to sample efficiently from log-concave distributions. What these, and other, works demonstrate is that certain optimization algorithms can be turned into sampling algorithms, and vice versa. What is lacking, however, is a generic reduction between these two tasks.

In the statistical literature, there are many results to the effect that, under suitable conditions, when data are sampled from some model in Θ, the maximum likelihood estimate in Θ asymptotically converges to this same model, and the posterior distribution asymptotically concentrates around it. Examples include the classical work of Le Cam [23]. Our hardness results are in a different setting—the data are arbitrary—but interestingly, require similar conditions on the prior. This is because our duplication technique gives the problem a statistical aspect: the final replicated data look rather like multiple draws from an underlying distribution supported on the initial data points.

3.2 Preliminaries and definitions

Let $\mathcal{X}$ be any data space. A *parameterized probability model* on $\mathcal{X}$ is a pair (p, Θ), where $p(\cdot \mid \theta)$ is a probability density over $\mathcal{X}$ for all $\theta \in \Theta$. We will be working with i.i.d. probability models, where for any sequence $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and any $\theta \in \Theta$,

$$p(X \mid \theta) = p(x_1, \ldots, x_n \mid \theta) = p(x_1 \mid \theta) \cdots p(x_n \mid \theta).$$

If we couple our probability model with a prior probability measure ν_0 over Θ, then the resulting triple (ν_0, p, Θ) is a *Bayesian parameterized probability model.* Let q_0 be a probability density corresponding to measure ν_0. The posterior density after observing X is then written as $q_X(\theta) \propto q_0(\theta) p(X \mid \theta)$ and we denote the corresponding measure as ν_X.

This notation conceals problem size: in reality, each input instance has some dimension m (the vocabulary size for documents, for instance) and requires parameters of corresponding dimensionality, in some $\Theta_m \subset \Theta$. We will suppress this dependence except where needed.

3.2.1 Maximum likelihood estimation

We formally define the maximum likelihood estimation problem as follows.

MAXIMUM LIKELIHOOD ESTIMATION: MLE-(p, Θ)
Input: A sequence of points $X \in \mathcal{X}^n$ and an accuracy parameter b in unary.
Output: A parameter $\theta \in \Theta$ satisfying

$$\log p(X \mid \theta) \geq \sup_{\theta' \in \Theta} \log p(X \mid \theta') - 1/b.$$

It might also be reasonable to ask for precision $1/2^b$. We adopt this particular formulation because it yields stronger hardness results.

3.2.2 Topic modeling

We will consider as a running example the problem of topic modeling. We follow the model of [4] where there is an unknown $V \times K$ topic matrix Ψ such that each column $\Psi^{(i)}$ is a distribution over a dictionary $[V]$, and there is a collection of D unknown, stochastically-generated distributions $\theta^{(1)}, \dots, \theta^{(D)}$ over the topics $[K]$. The standard choice of prior on $\theta^{(1)}, \dots, \theta^{(D)}$ is a symmetric Dirichlet(α) distribution. The generative process for a document d with words $w_1^{(d)}, w_2^{(d)}, \dots$ is

$$\theta^{(d)} \sim \text{Dirichlet}(\alpha), \qquad z_i \mid \theta^{(d)} \sim \text{Categorical}(\theta^{(d)}), \qquad w_i^{(d)} \mid z_i, \Psi^{(z_i)} \sim \text{Categorical}(\Psi^{(z_i)})$$

We observe the bags of words $X = [X^{(1)} | \cdots | X^{(D)}]$ where $X_i^{(d)} = |\{j : w_j^{(d)} = i\}|$. Since each document is generated independently and $\theta^{(1)}, \dots, \theta^{(D)}$ are assumed to be generated i.i.d., the likelihood of the corpus under a topic matrix Ψ is

$$p(X \mid \Psi) = \prod_{d=1}^{D} \mathbb{E}_{\theta^{(d)}} \left[p\left(X \mid \Psi, \theta^{(d)} \right) \right] = \prod_{d=1}^{D} \mathbb{E}_{\theta^{(d)}} \left[\prod_{i=1}^{V} \left(\sum_{k=1}^{K} \Psi_i^{(k)} \theta_k^{(d)} \right)^{X_i^{(d)}} \right]$$

How many bits does it take to approximate the maximum-likelihood Ψ? In the appendix, we show that for any discrete distribution $p = (p_1, \dots, p_\ell)$ and any $\epsilon > 0$, there is a rounded distribution $\widehat{p}$ that uses $\lceil \log_2(\ell/\epsilon) \rceil$ bits per entry and has $\widehat{p}_i \geq p_i(1-\epsilon)$. By applying this construction to each individual topic distribution, we get the following.

Lemma 3.2. *Consider any $V \times K$ topic distribution matrix Ψ. For any $\epsilon > 0$ and any integer m, there is a topic matrix $\widehat{\Psi}$ that uses $\lceil \log_2(mV/\epsilon) \rceil$ bits per entry, such that $\log p(x|\Psi) - \log p(x|\widehat{\Psi}) \leq \epsilon$ for all documents x of $\leq m$ words.*

Thus there exists a polynomially-sized solution to the problem of approximating the ML estimate of the topic modeling problem with a Dirichlet(α) prior on Θ, which we will refer to as TM-MLE(α). Arora et al. [4] demonstrated that TM-MLE(α) is NP-hard for

$\alpha = 1$. Their proof method works for any $\alpha > 0$; for completeness we present the following generalization of their result in the appendix.

Theorem 3.3. *[Implicit in [4]] We say a topic matrix Ψ is c-*smooth *for $c > 0$ if $\min_i \max_j \Psi_i^{(j)} \geq c$. Given $\alpha > 0$,* TM-MLE(α) *is NP-hard when $K = 2$, all the documents are restricted to have 2 words, and Ψ_{ML} is guaranteed to be $(1/V)$-smooth.*

The result given in the appendix is slightly more general in that it applies to all symmetric priors and not just the Dirichlet. However, to keep our examples concrete we will only refer to the NP-hardness of TM-MLE(α). Given this result, what can we say about the complexity of MAP estimation and sampling for topic modeling?

3.3 MAP estimation is as hard as ML estimation

In this section we give a generic reduction from ML estimation to MAP estimation. For a fixed data space $\mathcal{X}$, let (p, Θ) be a parameterized probability model and let ν_0 be a prior probability measure with an associated density q_0. Recall that we use the notation q_X to denote the posterior density given data X. We define the MAP estimation problem as follows.

MAXIMUM A POSTERIOR ESTIMATION: MAP-(p, q_0, Θ)
Input: A sequence of points $X \in \mathcal{X}^n$ and accuracy parameter b in unary.
Output: A parameter $\theta \in \Theta$ satisfying

$$\log q_X(\theta) \geq \sup_{\theta' \in \Theta} \log q_X(\theta') - 1/b.$$

For any instance $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $Z = (X^{(1)}, \ldots, X^{(k)})$ be a sequence consisting of k copies of X. The lemma below relates the MAP estimate for Z to the ML estimate for X.

Lemma 3.4. *Pick any $\delta > 0$ and any $\theta \in \Theta$ within δ of the optimal MAP solution for Z,* *that is,*

$$\log q_Z(\theta) \geq \sup_{\theta' \in \Theta} \log q_Z(\theta') - \delta.$$

Then the log-likelihood of any $\theta' \in \Theta$ can exceed that of θ by at most

$$\log p(X|\theta') - \log p(X|\theta) \leq \frac{1}{k}\left(\delta + \log q_0(\theta) - \log q_0(\theta')\right).$$

Lemma 3.4 is a promising start to a reduction from ML estimation to MAP estimation, but it requires the prior density $q(\cdot)$ to be bounded above and below, which is often not the case. Recall our example of topic modeling, where we are given a matrix bag of words $X \in \mathcal{Z}^{V\times K}$ and the ML goal is to find the topic matrix $\Psi \in \mathbb{R}^{V\times K}$ which maximizes the objective

$$\log p(X \mid \Psi) = \sum_{d=1}^{D} \log \mathbb{E}_{\theta^{(d)}}\left[\prod_{i=1}^{V}\left(\sum_{k=1}^{K} \Psi_i^{(k)}\theta_k^{(d)}\right)^{X_i^{(d)}}\right]$$

where $\theta^{(d)} \sim$Dirichlet(α). A common choice of prior for Ψ is to assume that the columns of Ψ are drawn i.i.d. from a symmetric Dirichlet(β) distribution. If we let q denote the density for the Dirichlet(β) distribution, then this prior density on Ψ can be written as

$$q_0(\Psi) = q(\Psi^{(1)})\, q(\Psi^{(2)}) \,\cdots\, q(\Psi^{(K)}).$$

We call the problem of maximizing the resulting posterior TM-MAP(α,β). For $\beta < 1$, the density q is not bounded from above, and TM-MAP(α,β) is consequently not well-defined: infinite a-posteriori scores can be achieved. Hence we will focus on the case $\beta \geq 1$. Here, however, q approaches 0 on the boundary of the simplex, which is problematic for the reduction because the ML solutions Ψ_{ML} of Theorem 3.3 contain topic distributions that are arbitrarily close to this boundary. Thus Lemma 3.4 cannot be straightforwardly applied using $\theta' = \Psi_{ML}$. Instead, we need to ensure that, for any data set, there is some intermediate Ψ that is far enough from the boundary to have non-negligible probability mass under q_0 but has high enough likelihood to be considered a good estimate of Ψ_{ML}.

In summary, we want to guarantee that there are good ML estimates with non-negligible weight under the prior density. The following two sections will help us formalize these notions.

3.3.1 Admissible distances

Given a likelihood function p, we define the *log-likelihood distance* between $\theta_1, \theta_2 \in \Theta$ as

$$d_p(\theta_1, \theta_2) \;=\; \sup_{x \in \mathcal{X}} |\log p(x \,|\, \theta_1) - \log p(x \,|\, \theta_2)|.$$

Note that for any data sequence $X \in \mathcal{X}^n$, we have

$$d_{p,X}(\theta_1, \theta_2) \;:=\; |\log p(X \,|\, \theta_1) - \log p(X \,|\, \theta_2)| \;\leq\; n\, d_p(\theta_1, \theta_2).$$

In our setting, where we are concerned with how close parameters are in terms of their log-likelihood, d_p is a natural distance to consider. However, it is often difficult to work with this distance directly. Indeed, when analyzing the behavior of a prior density over certain neighborhoods of a parameter space, there is often a much more convenient distance, such as an ℓ_p distance. In short, we need a way to relate log-likelihood distances to other parameter space distances that might be easier to work with.

Definition 3.5. *Given $\lambda \geq 1$ and $S \subset \Theta$, we say a distance $d(\cdot, \cdot)$ is (λ, S)*-admissible *if $d_p(\theta_1, \theta_2) \leq \lambda d(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in S$ such that $d(\theta_1, \theta_2) < 1/\lambda$. We call a λ satisfying this relationship an* admissibility constant*.*

Returning to our topic model example, define the max-norm distance between two topic matrices as

$$\|\Psi - \Phi\|_{max} = \max_{i,j} |\Psi_i^{(j)} - \Phi_i^{(j)}|$$

Then the following lemma demonstrates that max-norm distance is admissible over the set of smooth topic matrices, where smoothness was defined in Theorem 3.3.

Lemma 3.6. *Let $c, m > 0$ and suppose that Ψ and Φ are $V \times K$ c-smooth topic matrices such that $\|\Psi - \Phi\|_{max} \leq \min(c/m, 1/2)$. If $\alpha_0 = \sum \alpha_i$, then for any document x with length bounded by m,*

$$|\log p(x \mid \Psi) - \log p(x \mid \Phi)| \;\leq\; \|\Psi - \Phi\|_{max} \left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m} \right).$$

It is important to point out that although our discussion of topic modeling has assumed a symmetric Dirichlet distribution, the above lemma holds for non-symmetric Dirichlet distributions as well. Additionally, for the instances produced in Theorem 3.8, $K = n = 2$ and $c = 1/V$; thus the parenthesized term is polynomial in V.

Finally, for $\theta \in \Theta$ and $\epsilon > 0$, define the ball around θ of radius ϵ with respect to distance d as the set $B_d(\theta, \epsilon) = \{\theta' \in \Theta \,:\, d(\theta, \theta') < \epsilon\}$.

3.3.2 Promise problems

When constructing polynomial time reductions from a language L to a language L', the typical approach is to demonstrate the existence of a polynomial-time computable function $f : \Sigma^* \to \Sigma^*$ such that $x \in L$ if and only if $f(x) \in L'$. However, it is often the case that reductions only demonstrate the hardness of certain well-behaved subsets of languages. Such subsets are captured in the notion of *promise problems.* Given a function $\Pi : \Sigma^* \to \{0, 1\}$, known as the promise, and a language $L \subset \Sigma^*$, the promise problem Π-L is the problem of determining if $x \in L$ given input instances with $\Pi(x) = 1$.

3.3.3 The reduction

To turn Lemma 3.4 into a generic reduction, we need to assert that for any valid data sequence X and any $\epsilon > 0$, there is θ_ϵ whose log-likelihood is within ϵ of θ_{ML} such that $q_0(\theta_\epsilon)$ is bounded below from zero. Such a condition on θ_{ML} is implicitly a restriction on valid inputs X and therefore can be phrased as a promise.

Theorem 3.7. *Let m be any measure of the size of an input instance, and $\lambda(m)$ any function of this size. Let $S \subset \Theta$ be some subset of parameters, and d be a $(\lambda(m), S)$-admissible distance function. Suppose q_0 satisfies two properties:*

(i) it is bounded above by $2^{\mathrm{poly}(\lambda(m))}$ and

(ii) given $\epsilon > 0$ and $\theta \in S$, there exists $\theta_\epsilon \in B_d(\theta, \epsilon) \cap S$ such that $\log q_0(\theta_\epsilon) \geq -\mathrm{poly}(\lambda(m), 1/\epsilon)$.

If Π is the promise that $\theta_{ML} \in S$, then Π-MLE-$(p, \Theta) \leq_P$ MAP-(p, q_0, Θ), where the reduction is polynomial in the input length and $\lambda(m)$.

Proof. Let $X = (x_1, \ldots, x_n)$ and $b = 1^b$ be the input to MLE-(p, Θ) and let Z denote the sequence consisting of k copies of X. Our input to MAP-(p, q_0, Θ) will be the data sequence Z and the accuracy parameter b. Suppose that the output of this call is θ. L

Let $\epsilon = 1/(2bn\lambda(m))$ and take θ_ϵ to be the point satisfying $\theta_\epsilon \in S \cap B_d(\theta_{ML}, \epsilon)$ and $q_0(\theta_\epsilon) \geq 2^{-\mathrm{poly}(\lambda(m), 1/\epsilon)}$ whose existence is guaranteed by assumption (ii). By Lemma 3.4,

$$\begin{aligned}
\left|\log \frac{p(X \mid \theta_{ML})}{p(X \mid \theta)}\right| &\leq |\log p(X \mid \theta_{ML}) - \log p(X \mid \theta_\epsilon)| + |\log p(X \mid \theta_\epsilon) - \log p(X \mid \theta)| \\
&\leq n \cdot d_p(\theta_{ML}, \theta_\epsilon) + |\log p(X \mid \theta_\epsilon) - \log p(X \mid \theta)| \\
&\leq \lambda(m) n \cdot d(\theta_{ML}, \theta_\epsilon) + \frac{1}{k}\left(\frac{1}{b} + \log \frac{q_0(\theta)}{q_0(\theta_\epsilon)}\right) \\
&\leq \lambda(m) n \epsilon + \frac{1}{k}\left(\frac{1}{b} + \log \frac{2^{\mathrm{poly}(\lambda(m))}}{2^{-\mathrm{poly}(\lambda(m), 1/\epsilon)}}\right)
\end{aligned}$$

$$\leq \lambda(m)n\epsilon + \frac{1}{k}\left(\frac{1}{b} + \mathrm{poly}(\lambda(m), 1/\epsilon)\right).$$

By taking k to be a large enough polynomial in b, $\lambda(m)$, and n, we can guarantee that θ is within $1/b$ of the ML solution. □

Returning to topic modeling, let Π denote the promise that the data sequence has only 2 words per document and the ML solution is $1/V$-smooth. From Theorem 3.3, Π-TM-MLE(α) is NP-hard for any fixed $\alpha > 0$.

Now take C to be the admissibility constant from Lemma 3.6 with $c = 1/V$, $n = K = 2$. In the appendix, we show that when the prior q_o is Dirichlet(β) with $\beta \geq 1$, then for any $\epsilon > 0$ and any input instance, there exists a $1/V$-smooth Ψ_ϵ that is within ϵ of Ψ_{ML} in max-norm distance and satisfies $\log q_0(\Psi_\epsilon) \geq -\mathrm{poly}(C, 1/\epsilon)$. Theorem 3.7 immediately gives us the following.

Theorem 3.8. *For any fixed $\alpha > 0$ and $\beta \geq 1$,* TM-MAP(α, β) *is NP-hard.*

3.4 Approximate sampling is as hard as ML estimation

We now turn to giving a generic reduction from ML estimation to posterior sampling. As pointed out in Section 3.1, total variation distance might not be a suitable metric for approximate sampling in continuous domains. Thus we begin with a carefully chosen notion of approximate sampling. Afterwards, we give a reduction from ML estimation to approximate sampling and demonstrate how it applies to topic modeling.

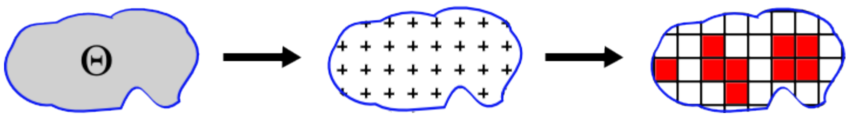


Figure 3.1. Rounding our samples produces a gridding of Θ. The resulting distribution is indistinguishable from the original distribution with respect to any union of grid boxes.

3.4.1 Notions of approximate sampling

Given measures μ and ν over a set Θ and a collection $\mathcal{B}$ of measurable subsets of Θ, define the *$\mathcal{B}$-variation distance* as

$$d_{\mathcal{B}}(\mu, \nu) = \sup_{B \in \mathcal{B}} |\mu(B) - \nu(B)|.$$

When $\mathcal{B}$ is the collection of all measurable subsets, this is total variation distance. For smaller collections, $d_{\mathcal{B}}$ may differ significantly from d_{TV} but is still a pseudometric.

What are minimal requirements on $\mathcal{B}$ to ensure that $d_{\mathcal{B}}$ is a meaningful probability distance? Suppose that Θ is equipped with a pseudometric $d(\cdot, \cdot)$; and, to avoid pathologies, assume (Θ, d) is separable (has a countable dense subset). Define $B_d(\theta, r) = \{\theta' \in \Theta : d(\theta, \theta') < r\}$. For $\epsilon > 0$ and $c \geq 1$, we say that a collection $\mathcal{B}$ is (d, c, ϵ)-*fine* if for every point $\theta \in \Theta$ there exists a $B \in \mathcal{B}$ such that $B_d(\theta, \epsilon) \subset B \subset B_d(\theta, c\epsilon)$. Intuitively, $\mathcal{B}$ captures the space Θ at a resolution of roughly ϵ.

For total variation distance, the supremum is taken over all measurable sets, which are closed under countable union and intersection. Likewise, we say $\mathcal{B}$ is a *standard* collection if it is closed under countable union. Note that if we have a (d, c, ϵ)-fine collection and consider its closure under countable union, the result remains (d, c, ϵ)-fine.

To understand the effect of choosing a family of sets $\mathcal{B}$, consider a simple example: suppose we sample from some distribution μ over Θ and then round the sample to r bits of precision. What is a suitable family $\mathcal{B}$? One option, illustrated in Figure 3.1, is to grid Θ with boxes of width $O(2^{-r})$, and let $\mathcal{B}$ be all unions of such boxes.

The following theorem generalizes this intuition and demonstrates the existence of standard (d, c, ϵ)-fine collections as well as the existence of perfect discretizations of arbitrary distributions.

Theorem 3.9. *Let ν be a distribution over a space Θ equipped with a pseudometric $d(\cdot, \cdot)$. For $\epsilon > 0$, suppose $\widehat{\Theta}$ is a countable ϵ-cover of Θ with respect to d. Then there exists a standard collection $\mathcal{B}$ of measurable subsets and a discrete measure $\widehat{\nu}$ over $\widehat{\Theta}$ such that*

(i) $\mathcal{B}$ is (d, c, ϵ)-fine for $c = 3$,

(ii) $d_{\mathcal{B}}(\widehat{\nu}, \nu) = 0$, and

(iii) for any discrete distribution $\widehat{\mu}$ over $\widehat{\Theta}$, $d_{\mathcal{B}}(\widehat{\mu}, \nu) = d_{TV}(\widehat{\mu}, \widehat{\nu})$.

Proof. For every $\widehat{\theta} \in \widehat{\Theta}$, define the inner Voronoi cell of $\widehat{\theta}$ to be

$$C^i(\widehat{\theta}) \;:=\; \{\theta \,:\, d(\theta, \widehat{\theta}) < d(\theta, \bar{\theta}) \;\; \forall \bar{\theta} \in \widehat{\Theta} \setminus \{\widehat{\theta}\}\}.$$

The Voronoi cell $C(\widehat{\theta})$ consists of $C^i(\widehat{\theta})$ as well as part of its boundary. To ensure that these cells are disjoint and cover all of Θ, we can order $\widehat{\Theta}$ and adopt the convention that the boundary occurring among any Voronoi cells belongs to the cell whose center comes earliest in the ordering.

Define $\mathcal{B}$ to be the union-closure of the set of Voronoi cells:

$$\mathcal{B} \;=\; \left\{ \cup_{\widehat{\theta} \in \mathcal{I}} C(\widehat{\theta}) \,:\, \mathcal{I} \subset \widehat{\Theta} \right\}.$$

By the countability of $\widehat{\Theta}$ we have that $\mathcal{B}$ is closed under countable union. To see that $\mathcal{B}$ is (d, c, ϵ)-fine we need to show that for every $\theta \in \Theta$, there exists a $B \in \mathcal{B}$ such that

$$B_d(\theta, \epsilon) \;\subset\; B \;\subset\; B_d(\theta, 3\epsilon).$$

Let B be the union of Voronoi cells that intersect $B_d(\theta, \epsilon)$. The first set inclusion follows immediately. To see the second set inclusion, note that because $\widehat{\Theta}$ is an ϵ-covering, $C(\widehat{\theta}) \subset B_d(\widehat{\theta}, \epsilon)$. If $C(\widehat{\theta}) \cap B_d(\theta, \epsilon) \neq \emptyset$, then we have $d(\widehat{\theta}, \theta) \leq 2\epsilon$. This implies that $C(\widehat{\theta}) \subset B_d(\theta, 3\epsilon)$. Thus, the union of such sets must also be contained in $B_d(\theta, 3\epsilon)$.

Now let $\widehat{\nu}$ denote the discrete distribution over $\widehat{\Theta}$ such that $\widehat{\nu}(\widehat{\theta}) = \nu(C(\widehat{\theta}))$. Then any $B \in \mathcal{B}$ is the countable union of such sets, so we have $\widehat{\nu}(B) = \nu(B)$, which implies $d_{\mathcal{B}}(\widehat{\nu}, \nu) = 0$.

Now consider $\widehat{\mu}$ to be any other discrete distribution over $\widehat{\Theta}$. For any $\delta > 0$, there is some $A_\delta \subset \widehat{\Theta}$ that achieves

$$|\widehat{\mu}(A_\delta) - \widehat{\nu}(A_\delta)| \;\geq\; d_{TV}(\widehat{\mu}, \widehat{\nu}) - \delta.$$

If $B = \cup_{\widehat{\theta} \in A_\delta} C(\widehat{\theta})$, then

$$d_{TV}(\widehat{\mu}, \widehat{\nu}) \;\leq\; |\widehat{\mu}(A) - \widehat{\nu}(A)| + \delta \;=\; |\widehat{\mu}(B) - \nu(B)| + \delta \;\leq\; d_{\mathcal{B}}(\widehat{\mu}, \nu) + \delta.$$

But because $d_{\mathcal{B}}$ is a pseudometric, we have

$$d_{\mathcal{B}}(\widehat{\mu}, \nu) \;\leq\; d_{\mathcal{B}}(\widehat{\mu}, \widehat{\nu}) + d_{\mathcal{B}}(\widehat{\nu}, \nu) \;=\; d_{\mathcal{B}}(\widehat{\mu}, \widehat{\nu}) \;=\; d_{TV}(\widehat{\mu}, \widehat{\nu}).$$

Since our choice of $\delta > 0$ was arbitrary, we can conclude $d_{TV}(\widehat{\mu}, \widehat{\nu}) = d_{\mathcal{B}}(\widehat{\mu}, \nu)$. □

Since c takes a constant value in Theorem 3.9, we will say a collection is (d, ϵ)-fine if it is (d, c, ϵ)-fine for some constant c. With these notions in hand, we are ready to give the definition of the approximate sampling problem.

> APPROXIMATE POSTERIOR SAMPLING: APPROX-SAMPLING-(p, ν_0, Θ)-d
> *Input*: A sequence of points $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$, accuracy parameter b in unary.
> *Output*: A parameter $\theta \in \Theta$ satisfying $\theta \sim \nu$ such that $d_{\mathcal{B}}(\nu, \nu_X) \leq 1/b$ where $\mathcal{B}$ is a standard $(d, 1/b)$-fine collection.

When can we guarantee that a θ from the above problem will be polynomially sized? If we take $\widehat{\Theta}$ to be a $1/b$-covering of Θ, then Theorem 3.9 guarantees the existence of a $(d, 1/b)$-fine collection $\mathcal{B}$ and discrete distribution ν over $\widehat{\Theta}$ such that $d_{\mathcal{B}}(\nu, \nu_X) = 0$. In the case where Θ is a bounded subset of $\mathbb{R}^m$ and d is an ℓ_p norm, for example, every element of $\widehat{\Theta}$ can be written using a polynomial number of bits. Thus, every draw from ν will be polynomially sized.

3.4.2 The reduction

Recall the definition of $d_{p,X}$ as $d_{p,X}(\theta_1, \theta_2) = |\log p(X|\theta_1) - \log p(X|\theta_2)|$ for a data sequence $X \in \mathcal{X}^n$ and $\theta_1, \theta_2 \in \Theta$. The following lemma tells us the rate at which the posterior of a duplicated data sequence concentrates around the ML solution.

Lemma 3.10. *Take any $\epsilon, \delta > 0$ and $X \in \mathcal{X}^n$. If Z is the sequence created by duplicating X k times for*

$$k \geq \frac{2}{\epsilon}\left(\log\left(\frac{1}{\delta} - 1\right) + \log\left(\frac{1 - \nu_0(B_{d_{p,X}}(\theta_{ML}, \epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ML}, \epsilon/2))}\right)\right)$$

then $\nu_Z(B_{d_{p,X}}(\theta_{ML}, \epsilon)) \geq 1 - \delta$.

With this in hand, we can state the main theorem of the section. We need to make similar considerations with respect to distances and promises that we did in Section 3.3.

Theorem 3.11. *Let m be any measure of the size of an input instance, and let $\lambda(m)$ be any function of this size. Let d be a distance function and $S \subset S' \subset \Theta$ be subsets satisfying*

(i) if $\theta \in S$ then $B_d(\theta, 1/\lambda(m)) \subset S'$ and

(ii) d is $(\lambda(m), S')$-admissible.

If Π is the promise that $B_{d_p}(\theta_{ML}, 1/\lambda(m)) \subset S$ and $\nu_0(B_d(\theta_{ML}, \epsilon)) \geq 2^{-\operatorname{poly}(\lambda(m), 1/\epsilon)}$ for all $\epsilon > 0$, then Π-MLE-$(p, \Theta) \leq_P$ APPROX-SAMPLING-(p, Θ, ν_0)-d under randomized reductions which are polynomial in the input size and $\lambda(m)$.

Proof. Let $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and b be input to Π-MLE-(p, Θ) and let $\delta > 0$. If Π is not true, then we can return anything and terminate.

Otherwise, let $\epsilon > 0$ and take $B_\epsilon = \{\theta : |\log p(X|\theta) - \log p(X|\theta_{ML})| < \epsilon\}$. By Lemma 3.10, we can duplicate the data

$$k(\epsilon, \delta) \;=\; \frac{2}{\epsilon} \log\left(\frac{1}{\delta} - 1\right) - \log \nu_0 \left(B_d \left(\theta_{ML}, \frac{\epsilon}{2n\lambda(m)}\right)\right)$$

times to ensure $\nu_Z(B_{d_{p,X}}(\theta_{ML}, \epsilon)) \geq 1 - \delta$ for $Z = (X^{(1)}, \ldots, X^{(k(\epsilon,\delta))})$.

If our accuracy parameter given to APPROX-SAMPLING-(p, Θ, ν_0)-d is b', the collection $\mathcal{B}$ our approximate distribution is measured against is a standard $(d, c, 1/b')$-fine collection. Thus, for every $\theta \in \Theta$, there exists a $B_\theta \in \mathcal{B}$ such that $B_d(\theta, 1/b') \subset B_\theta \subset B_d(\theta, c/b')$. Since $\mathcal{B}$ is standard, we also have the set

$$B \;=\; \bigcup_{\theta \in B_{d_{p,X}}(\theta_{ML}, \epsilon)} B_\theta$$

is in $\mathcal{B}$. Therefore, if $\widehat{\nu}$ satisfies $d_{\mathcal{B}}(\widehat{\nu}, \nu_Z) \leq \delta$, then

$$\widehat{\nu}(B) \;\geq\; \nu(B) - \delta \;\geq\; 1 - 2\delta.$$

From this we know if $\theta \sim \widehat{\nu}$, then $\theta \in B$ with probability $1 - 2\delta$. Let us condition on this occurring. Then there exists $\theta' \in B_{d_{p,X}}(\theta_{ML}, \epsilon)$ such that $d(\theta, \theta') \leq c/b'$. For $\epsilon < \frac{1}{n\lambda(m)}$ and $b' \geq c/\epsilon$, we have $\theta' \in S$ and $\theta \in S'$ and

$$\begin{aligned}
\left|\log \frac{p(X|\theta)}{p(X|\theta_{ML})}\right| \;&=\; |\log p(X|\theta) - \log p(X|\theta')| + |\log p(X|\theta') - \log p(X|\theta_{ML})| \\
&\leq\; nd_p(\theta, \theta') + |\log p(X|\theta') - \log p(X|\theta_{ML})| \\
&\leq\; n\lambda(m) d(\theta, \theta') + |\log p(X|\theta') - \log p(X|\theta_{ML})| \\
&\leq\; \epsilon(n\lambda(m) + 1)
\end{aligned}$$

By taking $\epsilon = 1/(n\lambda(m) + 1)$, $k = k(\epsilon, \delta)$, and $b' = cb/\epsilon$, then we know that if our input to APPROX-SAMPLING-(p, Θ, ν_0)-d is the data sequence $Z = (X^{(1)}, \ldots, X^{(k)})$ and the accuracy parameter b', then with probability at least $1 - 2\delta$ the draw from APPROX-SAMPLING-(p, Θ, ν_0)-d is within $1/b$ of θ_{ML}. □

To see how this applies to our topic modeling scenario, recall that our posterior was formed by considering the likelihood in TM-MLE(α) and placing a Dirichlet(β) prior on each of the columns of Ψ. We call the problem of sampling from this distribution TM-APPROX-SAMPLING(α, β).

Notice that Theorem 3.11 only requires a lower bound on the probability of neighborhoods of ML solutions and not any type of upper bound as in Theorem 3.7. Therefore, we do not need to place the same lower bound on β as in the MAP estimation reduction. In particular, we prove the following in the appendix.

Theorem 3.12. *For any fixed $\alpha, \beta > 0$,* TM-APPROX-SAMPLING(α, β) *is NP-hard.*

3.5 Application: mixtures of Gaussians

Recall the maximum likelihood estimation problem for mixtures of k spherical Gaussians introduced in Chapter 2.

> MLE FOR MIXTURES OF k SPHERICAL GAUSSIANS WITH SAME VARIANCE:
> MLE-MOG-SV(k)
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; unary parameter b.
> *Output*: A mixture of k spherical Gaussians with the same variance, $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, whose log-likelihood
>
> $$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma^2) \right).$$
>
> is within an additive factor $1/b$ of optimal.

As we saw, this problem is NP-hard.

We now turn to the corresponding Bayesian problems for mixtures of Gaussians. We will consider common conjugate priors on our mixing weights, means, and variance. In particular, we will place a symmetric Dirichlet(γ) prior on the mixing weights and a Normal-Inverse-Gamma($\alpha, \beta, \mu_0, n_0$) prior on the means and variance, wherein the variance σ^2 is first drawn from an inverse gamma distribution – IG(α, β) – and the means are drawn i.i.d. from a normal distribution – $N(\mu_0, \sigma^2/n_0 I_d)$. The full generative process can be spelled out as follows.

$$(\pi_1, \ldots, \pi_k) \sim Dir(\gamma) \qquad \mu_j \,|\, \sigma^2 \sim N(\mu_0, \sigma^2/n_0 I_d)$$

$$\sigma^2 \sim IG(\alpha, \beta) \qquad x_i | \boldsymbol{\mu}, \boldsymbol{\pi}, \sigma^2 \sim \sum_{i=1}^{k} \pi_i \, N(\mu_i, \sigma^2 I_d)$$

For a fixed set of hyper-parameters $\omega = (\alpha, \beta, \gamma, \mu_0, n_0)$, let us call the corresponding MAP estimation problem MAP-MOGS(k, ω) and the approximate sampling problem APPROX-SAMPLING-MOGS(k, ω). We will show that both of these problems are hard when $k = 2$.

As in the topic modeling setting, we cannot simply start with a reduction from MLE-MOGS-SV(k). We will need a well-behaved promise problem version of this problem.

Theorem 3.13. *Let Π be the promise that there exists a low-order polynomial $\rho(\cdot, \cdot, \cdot)$ such that if $\theta_{ML} = (\boldsymbol{\mu}^*, \boldsymbol{\pi}^*, \sigma^*)$ is an optimal maximum likelihood solution and $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ satisfies $d_p(\theta_{ML}, \theta) < 1$, then*

(i) $\|\mu_j\| \leq \rho(n, d, k)$ for all j,

(ii) $\sigma^2 \geq 1/\rho(n, d, k)$,

(iii) $\pi_j > 0$ for all j, and

(iv) $\pi_j^* \geq 1/\rho(n, d, k)$ for all j.

*Then Π-*MLE-MOGS-SV(k) *is NP-hard for $k \geq 2$.*

The proof of Theorem 3.13 is done in the exact same way as the proofs in Chapter 2 and is deferred to the appendix.

As before, we also need a suitable distance to dominate likelihood distance. We will consider the following distance between two parameters $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\sigma})$:

$$d(\theta, \hat{\theta}) \; = \; \max\left\{\|\mu_i - \hat{\mu}_i\|^2, |\log \pi_i - \log \hat{\pi}_i|, |\sigma^2 - \hat{\sigma}^2|\right\}.$$

The following lemma, whose proof appears in the appendix, shows that this distance does indeed dominate likelihood distance for well-behaved parameters.

Lemma 3.14. *Let $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\sigma})$ be two parameter vectors satisfying $\pi_j, \hat{\pi}_j > 0$ for all j. Then $d_p(\theta, \hat{\theta}) \leq d(\theta, \hat{\theta}) \operatorname{poly}(1/\sigma_i^2, 1/\hat{\sigma}_i^2, \|\mu_i\|^2, \|\hat{\mu}_i\|^2)$.*

Next, we give bounds on the prior density.

Lemma 3.15. *Let q and ν be the prior density and measure, respectively, for the Bayesian mixture of two spherical Gaussians generative model with parameters $\alpha, \beta, \gamma, \mu_0, n_0$. For any $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and any $\epsilon > 0$, we have*

$$\log q(\theta) \; \geq \; -\operatorname{poly}(1/\pi_i, 1/\sigma, |\mu_i\|, d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$$

and

$$\log \nu(B_d(\theta, \epsilon) \; \geq \; -\operatorname{poly}(1/\pi_i, 1/\sigma, |\mu_i\|, d, 1/\epsilon, n_0, \alpha, \beta, \gamma, \|\mu_0\|).$$

Further, if $\gamma \geq 1$, we have

$$\log q(\theta) \; \leq \; \operatorname{poly}(d, n_0, \alpha, \beta, \gamma, \|\mu_0\|).$$

The proof of Lemma 3.15, which is deferred to the appendix, boils down to simply separately bounding the Dirichlet and Normal-Inverse-Gamma distributions.

Given the above, we are now ready to show that MAP estimation and approximate posterior sampling are NP-hard in this setting. We start with MAP estimation.

Theorem 3.16. *Let $\omega = (\alpha, \beta, \gamma, \mu_0, n_0)$ for $\alpha, \beta, n_0 > 0$, $\gamma \geq 1$, and $\mu_0 \in \mathbb{R}^d$. Then* $\textsc{map-mogs}(2, \omega)$ *is NP-hard.*

Proof. We will reduce from Π-$\textsc{mle-mogs-sv}(2)$. Let q denote the prior density and let

$$S = \{(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma) \, : \, \sigma^2 \geq 1/\rho(n, d, k), \max_i \|\mu_i\|^2 \leq \rho(n, d, k), \min_i \pi_i > 1/\rho(n, d, k)\}.$$

Then we have the following.

(i) $\log q(\theta) \leq \mathrm{poly}(d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$ for any parameter $\theta \in \Theta$ (Lemma 3.15).

(ii) $\log q(\theta) \geq -\,\mathrm{poly}(n, d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$ for any parameter $\theta \in S$ (Lemma 3.15).

(iii) d is $(\mathrm{poly}(n, d), S)$-admissible (Lemma 3.14).

(iv) Π guarantees that $\theta_{ML} \in S$.

Given the above, Theorem 3.7 implies that $\textsc{map-mogs}(2, \omega)$ is NP-hard. □

We now turn to showing that approximate posterior sampling.

Theorem 3.17. *Let $\omega = (\alpha, \beta, \gamma, \mu_0, n_0)$ for $\alpha, \beta, \gamma, n_0 > 0$ and $\mu_0 \in \mathbb{R}^d$. Then* $\textsc{approx-sampling-mogs}(2, \omega)$ *is NP-hard.*

Proof. We again reduce from Π-$\textsc{mle-mogs-sv}(k)$ for $k = 2$. Since k is a constant, we may take the polynomial ρ from Theorem 3.13 to only have two free arguments. In order to apply Theorem 3.11, let $\rho(\cdot, \cdot)$ be the polynomial from Theorem 3.13 and let

$$S = \{(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma) \, : \, \sigma^2 \geq 1/\rho(n, d), \max_i \|\mu_i\|^2 \leq \rho(n, d), \min_i \pi_i > 0\}$$

$$S' = \{(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma) \, : \, \sigma^2 \geq 1/2\rho(n, d), \max_i \|\mu_i\|^2 \leq 2\rho(n, d), \min_i \pi_i > 0\}$$

$$S^* = \{(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma) \,:\, \sigma^2 \geq 1/\rho(n,d), \max_i \|\mu_i\|^2 \leq \rho(n,d), \min_i \pi_i > 1/\rho(n,d)\}$$

Then we have the following.

(i) If $\theta \in S$ then $B_d(\theta, 1/2\rho(n,d)) \subset S'$ (definition of distance d).

(ii) d is $(\mathrm{poly}(n,d), S')$-admissible (Lemma 3.14).

(iii) Π guarantees that $B_{d_p}(\theta_{ML}, 1) \subset S$ and $\theta_{ML} \in S^*$.

(iv) From (iii), $\log \nu_0(B_d(\theta_{ML}, \epsilon)) \geq -\mathrm{poly}(n, d, 1/\epsilon)$ for all $\epsilon > 0$ (Lemma 3.15).

Putting the above together, Theorem 3.11 implies that APPROX-SAMPLING-MOGS$(2, \omega)$ is NP-hard. □

Chapter 3 contains material that is currently being prepared for submission for publication of the material. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

Part II

Markov chains and mixing rates

Chapter 4

Markov chain preliminaries

4.1 A sampling problem

Consider the Bayesian approach to statistical modeling. We have some *prior distribution* q over a space of parameters Θ, and there is some unobserved random variable θ drawn according to Q. We then observe data points $x_1, \ldots, x_n$ distributed independently and identically according to a distribution parameterized by θ: P_θ. Given this setup, the posterior distribution is given by

$$Q_n(\theta) := \Pr(\theta \,|\, x_1, \ldots, x_n) = \frac{\Pr(\theta)\Pr(x_1, \ldots, x_n \,|\, \theta)}{\Pr(x_1, \ldots, x_n)} = \frac{1}{Z} Q(\theta) \prod_{i=1}^{n} P_\theta(x_i)$$

where Z is a normalizing constant that makes the distribution integrate to 1.

The posterior distribution is a central focus of Bayesian statistics, and many statistical inference tasks are done by computing expectations with respect to it. Unfortunately, the normalizing constant Z is difficult to compute in general [22], which makes working directly with the posterior difficult.

To circumvent this difficulty, practitioners often resort to Monte Carlo methods [42]. To illustrate these approaches, suppose that our goal is to compute the expected value of $f : \Theta \to \mathbb{R}$ under the posterior distribution. That is, we wish to compute $\mathbb{E}_{\theta \sim Q_n}[f(\theta)]$.

If we can draw samples $\theta_1, \ldots, \theta_m$ from some distribution $\widehat{Q}_n$ which is close to Q_n, then we can approximate the desired expectation via the empirical average

$$\frac{1}{m}(f(\theta_1) + \cdots + f(\theta_m)).$$

Thus, the problem of approximately computing expectations can be reduced to an approximate sampling problem. This leads us to the obvious question: how do we approximately sample from an intractable distribution?

4.2 Markov chains and mixing rates

A *Markov chain* is a stochastic process $(X_0)_{t=0}^{\infty}$ taking values in some state space Ω and satisfying the relation

$$\Pr(X_{t+1} \in A \,|\, X_0, X_1, \ldots, X_t) \;=\; \Pr(X_{t+1} \in A \,|\, X_t).$$

To simplify our setting, we will assume that the space Ω is large but finite. Under this assumption, we can consider the transition probabilities as a matrix indexed by the elements of Ω such that

$$Q(x, y) \;=\; \Pr(X_{t+1} = y \,|\, X_t = x).$$

Using this notation, it is not difficult to see that the k-fold product Q^k represents the k-step transition probabilities

$$Q^k(x, y) \;=\; \Pr(X_{t+k} = y \,|\, X_t = x).$$

Suppose that π is a probability vector indexed by the elements of Ω. The distribution of X_t, given that $X_0 \sim \pi$ can be succinctly written as

$$\Pr(X_t = x \,|\, X_0 \sim \pi) \;=\; \pi Q^t.$$

A distribution π is a *stationary distribution* of Q if $\pi \;=\; \pi Q$. To check that a distribution π is stationary with respect to Q, it is sufficient, but not necessary, to establish *reversibility*:

$$\pi(x)Q(x,y) \;=\; \pi(y)Q(y,x)$$

for all $x, y \in \Omega$. To see this, suppose this condition holds, then

$$(\pi Q)(x) \;=\; \sum_{y\in\Omega} \pi(y)Q(y,x) \;=\; \sum_{y\in\Omega} \pi(x)Q(x,y) \;=\; \pi(x)$$

where the last equality follows from the fact that $\sum_y Q(x,y) = 1$.

We say Q is *irreducible* if, for all $x, y \in \Omega$, there exists an integer $t > 0$ such that $Q^t(x,y) > 0$. It is aperiodic if $\gcd(\{t \,:\, Q^t(x,y) > 0\}) = 1$.

Given the above, the following is a fundamental result of Markov chain theory. It can be found for example, in [66].

Theorem 4.1 (Theorem 4.9 of [66])**.** *Suppose $(X_t)_{t=0}^{\infty}$ is a Markov chain with irreducible and aperiodic transition matrix Q. Then*

(i) Q has a unique stationary distribution π and

(ii) the distribution of X_t converges to π, regardless of initial distribution.

Theorem 4.1 is the basis of Markov chain Monte Carlo (MCMC) methods, a Monte Carlo technique in which samples are drawn from a Markov chain whose stationary distribution is the desired distribution.

Unfortunately, the convergence result of Theorem 4.1 only holds in the limit. Thus, to get practical guarantees, we need a notion of distribution approximation. A common choice of distribution distance in the Markov chain literature is *total variation distance*: given two probability measures μ, ν over Ω, their total variation distance is

$$\|\mu - \nu\|_{TV} \;:=\; \sup_{A \subset \Omega} |\mu(A) - \nu(A)| \;=\; \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|$$

where the supremum is taken over all measurable subsets of Ω. Total variation distance is a very powerful notion of distance, as we have the relationship

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sup_{f : \Omega \to [-1,1]} \mathbb{E}_{\omega \sim \mu}[f(\omega)] - \mathbb{E}_{\omega \sim \nu}[f(\omega)].$$

where the supremum is taken over all measurable functions. Thus, when our ultimate goal is to compute estimates of expectations, total variation is a natural distance to consider.

The *mixing rate* of a Markov chain Q with unique stationary distribution π is the function $\tau : (0,1) \to \mathbb{N}$ satisfying

$$\tau(\epsilon) \;:=\; \min \left\{ t \;:\; \max_{x \in \Omega} \|Q^t(x, \cdot) - \pi\|_{TV} < \epsilon \right\}.$$

The mixing rate is then roughly the number of steps needed before samples from a Markov chain look as though they are distributed according to the stationary distribution, up to tolerance ϵ. The quantity $\tau_{mix} = \tau(1/4)$ is sometimes also referred to as the mixing rate.

4.3 The Gibbs sampler

How do we construct Markov chains with a desired stationary distribution? There are several generic techniques to address this question. Of these, *Gibbs sampling* [43] is one of the most enduring and popular.

```
Initialize $X_0 \in \Omega^n$
For $t = 1, 2, \ldots$:
    • Pick a coordinate $i$
    • Sample $x \sim P_i(\cdot \,|\, X_{t-1}(-i))$
    • Set $X_t(-i) = X_{t-1}(-i)$ and $X_t(i) = x$
```

Figure 4.1. The generic Gibbs sampling algorithm.

Suppose our state space can be written as an n-fold product Ω^n and our desired stationary distribution can be written as the joint distribution $P(x_1, \ldots, x_n)$ for $x \in \Omega^n$. Moreover, suppose that it is possible to sample from the coordinate conditional distributions:

$$P_i(x_i \,|\, x_{-i}) \;=\; P_i(x_i \,|\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

where the notation x_{-i} denotes x with the i-th coordinate position removed. The Gibbs sampler, displayed in Figure 4.1, is a Markov chain whose transitions are governed by these coordinate conditional distributions.

Once the coordinate conditional distributions are fixed, the only the specification that needs to be made for the Gibbs sampler is the way the next coordinate is chosen. Two common variations are *random scan*, in which the next coordinate is chosen uniformly at random, and *systematic scan*, in which the coordinates are cycled through in a deterministic ordering.

Regardless of which ordering is chosen, the distribution of the Gibbs sampler is guaranteed to converge to $P(\cdot)$ so long as the coordinate distributions $P_i(x_i \,|\, x_{-i})$ are lower-bounded everywhere. To see this, suppose for simplicity that we are using the random scan Gibbs sampler. First notice that $P_i(x_i \,|\, x_{-i})$ being lower-bounded everywhere implies that the Gibbs sampler is irreducible and aperiodic. Then see that the Gibbs sampler is reversible with respect to $P(\cdot)$, since for $x, y \in \Omega^n$ differing at a single position i

$$\begin{aligned}
P(x)\Pr(X_{t+1} = y \,|\, X_t = x) &= P(x)\Pr(\text{choose coordinate } i)P_i(y_i \,|\, x_{-i}) \\
&= P_i(x_i \,|\, x_{-i})P(x_{-i})\Pr(\text{choose coordinate } i)P_i(y_i \,|\, x_{-i}) \\
&= P(y)\Pr(\text{choose coordinate } i)P_i(x_i \,|\, x_{-i}) \\
&= P(y)\Pr(X_{t+1} = x \,|\, X_t = y)
\end{aligned}$$

Moreover, for $x, y \in \Omega^n$ differing at more than a single position,

$$P(x)\Pr(X_{t+1} = y \,|\, X_t = x) \;=\; 0 \;=\; P(y)\Pr(X_{t+1} = x \,|\, X_t = y).$$

Thus, the Gibbs sampler is reversible with respect to $P(\cdot)$. By reversibility and Theorem 4.1, the Gibbs sampler converges in distribution to $P(\cdot)$.

For many distributions that arise in Bayesian statistics, it is easy to sample from the coordinate conditional distributions, making the Gibbs sampler relatively easy to implement. However, as we shall see, the mixing rate of the Gibbs sampler depends heavily on the particular distribution under consideration. Thus, despite its appealing simplicity, the Gibbs sampler may not always give us accurate samples within a reasonable running time.

4.4 Bounding the mixing rate

In this thesis, we will be concerned with bounding mixing rates. Here, we examine two mathematical tools for doing so.

4.4.1 Lower bounds via conductance

Let Q be a Markov chain over a state space Ω with stationary distribution π, the *conductance* of $S \subset \Omega$ is

$$\Phi(S) := \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x) Q(x, y)$$

and the conductance of Q, denoted by Φ^*, is the minimum conductance of any set S with $\pi(S) \leq 1/2$. Intuitively, conductance measures how easily a Markov chain can transition out of low probability subsets of the state space. Thus, we might expect that a Markov chain with low conductance should take longer to mix.

This intuition is made precise by the following theorem, which relates the mixing rate and conductance of a Markov chain and has appeared in several forms throughout the Markov chain literature, see for example [82, 40]. We state here the form presented in [66].

Theorem 4.2 (Theorem 7.8 of [66])**.** *For any aperiodic, irreducible Markov chain with conductance Φ^*,*

$$\tau_{mix} \;\geq\; \frac{1}{4\Phi^*}.$$

Because Φ^* is the set corresponding to the minimum conductance, Theorem 4.2 implies that *any* set S satisfying $\pi(S) \leq 1/2$ provides the lower bound

$$\tau_{mix} \;\geq\; \frac{1}{4\Phi(S)}.$$

4.4.2 Upper bounds via coupling

Let μ and ν be probability measures over Ω. A pair of random variables (X, Y) is a *coupling* of μ and ν if

$$\Pr(X \in A) \;=\; \mu(A)$$

$$\Pr(Y \in B) \;=\; \nu(B)$$

for all measurable sets A and B.

Couplings are convenient probabilistic tools for bounding distances between measures. The following lemma, whose proof can be found in [1], tells us that not only does any coupling provide an upper bound on the total variation distance between measures but also that there exists a coupling that achieves this bound.

Lemma 4.3 (Lemma 3.6 of [1]). *Let μ and ν be probability measures.*

(a) For any coupling (X, Y) of μ and ν, $\|\mu - \nu\|_{TV} \leq Pr(X \neq Y)$.

(b) There exists a coupling (X, Y) satisfying $\|\mu - \nu\|_{TV} = Pr(X \neq Y)$.

It can be quite cumbersome to work with couplings involving entire stochastic processes. It is often more convenient to restrict our attention to the class of Markovian couplings. A *Markovian coupling* of a Markov chain over Ω with transition matrix Q is a Markov chain (X_t, Y_t) over $\Omega \times \Omega$ whose transitions satisfy

$$\begin{aligned}
\Pr(X_{t+1} = x' \mid X_t = x, Y_t = y) \;&=\; Q(x, x'), \\
\Pr(Y_{t+1} = y' \mid X_t = x, Y_t = y) \;&=\; Q(y, y').
\end{aligned}$$

The following lemma relates Markovian couplings to mixing times. It dates back at least to Aldous [1] and can be found in the form we present, for example, in [60].

Lemma 4.4 (Lemma 4.1 of [60]). *Let (X_t, Y_t) be a Markovian coupling for Markov chain Z_t such that there exists a function $\tau_{couple} : (0, 1) \to \mathbb{N}$ satisfying*

$$\Pr(X_{\tau_{couple}(\epsilon)} \neq Y_{\tau_{couple}(\epsilon)} \mid X_0 = x, Y_0 = y) \;\leq\; \epsilon$$

for all $x, y \in \Omega$ and $\epsilon > 0$. Then the mixing rate for Z_t satisfies $\tau(\epsilon) \leq \tau_{couple}(\epsilon)$.

Thus, to find an upper bound on the mixing rate of a Markov chain, it suffices to construct an appropriate Markovian coupling.

Chapter 5

Lower bounds on the Gibbs sampler over mixtures of Gaussians

We saw in Chapter 3 that approximately sampling from the posterior distribution of a Bayesian Gaussian mixture model is NP-hard. Thus, for general data sets, we cannot hope for efficient algorithms for this problem. However, these results do not tell us anything about the complexity of this problem when our data is well-behaved.

In this chapter, we investigate a commonly used algorithm, the Gibbs sampler, for approximately sampling from the posterior of a Bayesian Gaussian mixture model. We will see that the Gibbs sampler can take a very long time to converge, even when the data looks as though it were actually generated by the model.

5.1 Mixture models and Gibbs sampling

Although our results for the Gibbs sampler will pertain specifically to mixtures of Gaussians, it will be instructive to first look at a broader class of mixture models: mixtures of exponential families of distributions.

5.1.1 The generative model

For a mixture model of k components, we assume that our mixing weights $(w_1, \ldots, w_k)$ are drawn from a symmetric k-dimensional Dirichlet distribution with a single parameter $\alpha > 0$. This is a distribution over the k-simplex,

$$\Delta_k = \left\{ (w_1, \ldots, w_k) \in \mathbb{R}^k \,\middle|\, \sum_{i=1}^{k} w_i = 1,\ w_i \geq 0 \text{ for all i} \right\},$$

and has probability density function

$$D_\alpha(w_1, \cdots, w_k) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{i=1}^{k} w_i^{\alpha-1}.$$

Here, $\Gamma(\cdot)$ is the gamma function. The parameters $\theta_i \in \Theta$ are i.i.d. draws from the prior distribution parameterized by some vector $\beta \in \mathbb{R}^s$. Call this distribution $\mathcal{Q}(\beta)$ and its probability density function $Q_\beta : \Theta \to \mathbb{R}$. The label, or assignment, z_i for each data point is drawn from k-dimensional categorical distribution Categorical$(w_1, \ldots, w_k)$ which gives probability mass w_i to item i. Finally, the point x_i is drawn from the distribution parameterized by θ_{z_i}. Call this $\mathcal{P}(\theta_{z_i})$ and its probability density function $P_{\theta_{z_i}}$. The generative process can be summarized as the following.

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \theta_1, \ldots, \theta_k &\sim \mathcal{Q}(\beta) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim \mathcal{P}(\theta_{z_i}) \end{aligned} \tag{5.1}$$

Suppose that we produce a sequence $x = (x_1, \ldots, x_n)$ from the above generative process. Then the joint distribution of all quantities is given as

$$\Pr(x, z, \theta, w) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{j=1}^{k} w_j^{\alpha-1} Q_\beta(\theta_j) \prod_{i=1}^{n} (w_{z_i} P_{\theta_{z_i}}(x_i)).$$

We denote by $C_j(z)$ the set of indices i for which $z_i = j$ and by $n(z)$ the vector whose jth element is $|C_j(z)|$. Here we think about $C_j(z)$ as being the jth ‘cluster.’

```
Initialize $z_1, \ldots, z_n \in \{1, \ldots, k\}$
For $t = 1, 2, \ldots$:
    • Choose $i$ u.a.r. from $\{1, \ldots, n\}$
    • Update $z_i$ according to $\Pr(z_i = j \mid z_{-i}, x_1, \ldots, x_n)$
```

Figure 5.1. The collapsed Gibbs sampler.

For a subset S of $\{1, \ldots, n\}$, we let $P_\theta(S)$ denote the probability of S under the specific model $\theta \in \Theta$:

$$P_\theta(S) := \prod_{i \in S} P_\theta(x_i)$$

and let $q(S)$ denote the probability of S given $\theta \sim \mathcal{Q}(\beta)$:

$$q(S) := \int_\Theta Q_\beta(\theta) P_\theta(S)\, d\theta.$$

In this chapter, we are interested in the posterior probability of a labeling z given a data sequence x:

$$\Pr(z|x) \propto \prod_{j=1}^{k} \left(\frac{\Gamma(n_j(z) + \alpha)}{\Gamma(\alpha)} q(C_j(z)) \right). \tag{5.2}$$

Denote $\Pr(z|x)$ by $\pi(z)$. Even when q is computable in closed form, there are no known exact methods for computing the normalizing factor of π. Thus, we turn to a Markov chain which has π as its stationary distribution.

5.1.2 The collapsed Gibbs sampler

The collapsed Gibbs sampler, shown in Figure 5.1, is a Markov chain designed to sample from $\pi(z) = \Pr(z|x)$. As discussed in Chapter 4, this Markov chain does indeed converge to π. However, to efficiently implement the collapsed Gibbs sampler, we still need to compute $\Pr(z_i = j \mid z_{-i}, x)$. To this end, let S be a subset of indices, i be an index, and define

$$\Delta(S, i) \ := \ \frac{q(S \cup \{i\})}{q(S \setminus \{i\})}.$$

With this notation, we can state the following lemma.

Lemma 5.1. $\Pr(z_i = j \mid z_{-i}, x)$ *is proportional to* $(\alpha + n_j(z_{-i}))\Delta(C_j(z), i)$.

The proof of Lemma 5.1 is deferred to the appendix.

5.2 Markov chains and equivalence classes

Identifiability makes it difficult to analyze the mixing time of P. If σ is a permutation over $\{1, \ldots, k\}$, then z and $\sigma(z) = (\sigma(z_1), \ldots, \sigma(z_n))$ contain the same information for most applications. In general, we are only really interested in the clustering of the points, not the specific number assigned to each cluster. However, the collapsed Gibbs sampler views z and $\sigma(z)$ as separate states. Thus, mixing results proved over the labeling space may not hold true for the space we care about. We will now see how to factor out this extraneous information by a suitable projection.

5.2.1 Equivalence classes of Markov chains

Consider the following setting: we have a state space Ω and an equivalence relation $\sim$ on Ω. Let $(X_t)_{t=1}^{\infty}$ be a Markov chain and consider the sequence over the equivalence classes $([X_t])_{t=1}^{\infty}$. Under what conditions is this a Markov chain? The following lemma, which may be found in [66], answers this question.

Lemma 5.2 (Lemma 2.5 of [66])**.** *Let* $(X_i)_{i=1}^{\infty}$ *be a Markov chain with state space* Ω *and transition matrix* P *and let* $\sim$ *be an equivalence relation over* Ω *with equivalence classes* $\Omega^{\sharp} = \{[x] : x \in \Omega\}$. *Assume* P *satisfies* $P(x, [y]) = P(x', [y])$ *for all* $x \sim x'$, *where* $P(x, [y]) := \sum_{y' \sim y} P(x, y)$. *Then* $([X_i])_{i=1}^{\infty}$ *is a Markov chain with state space* $\Omega^{\sharp}$ *and transition function* $P^{\sharp}([x], [y]) = P(x, [y])$.

```
Initialize a clustering $\mathbb{C} \in \Omega_{\leq k}(n)$
For $t = 1, 2, \ldots$:
    • Choose $i$ u.a.r. from $\{1, \ldots, n\}$
    • Move $i$ to $S \in \mathbb{C}$ with probability proportional to $(\alpha + |S \setminus \{i\}|)\Delta(S, i)$
    • Move $i$ to own set with probability proportional to $(k - |\mathbb{C}|) \cdot \alpha \cdot q(\{i\})$[1]
```

Figure 5.2. The projected Gibbs sampler.

The following lemma establishes the form of the stationary distribution for $P^\sharp$.

Lemma 5.3. *Let P, $P^\sharp$, Ω, $\Omega^\sharp$, and $\sim$ be as in Lemma 5.2. If P is reversible with respect to π, then $P^\sharp$ is reversible with respect to $\pi^\sharp([x]) = \pi([x]) := \sum_{x' \sim x} \pi(x)$.*

Proof. Let $x, y \in \Omega$ be given.

$$\pi^\sharp([x])P^\sharp([x],[y]) = \pi([x])P^\sharp([x],[y]) = \sum_{x' \sim x, y' \sim y} \pi(x')P(x',y')$$
$$= \sum_{x' \sim x, y' \sim y} \pi(y')P(y',x') = \pi([y])P^\sharp([y],[x]) = \pi^\sharp([y])P^\sharp([y],[x]). \quad \square$$

5.2.2 Induced clusterings

Consider the equivalence relation $\sim$ over labelings such that $z \sim z'$ if there exists a permutation σ s.t. $\sigma(z) = z'$. Let P denote the Gibbs sampler from Figure 5.1. What does the corresponding Markov chain over the equivalence classes, $P^\sharp$, look like?

Equation (5.2) tells us that $z', z'' \in [z]$ have the same probability mass under π. Thus, one way to describe $P^\sharp$ is that if the current state is $[z]$, it chooses any labeling $z' \in [z]$, moves to a neighboring labeling z'' according to P, and sets the new state to be $[z'']$.

[1] This is ambiguous if i is already its own cluster. In this case, the probability we keep i as its own set is proportional to $(k - |\mathbb{C}| + 1) \cdot \alpha \cdot q(\{i\})$.

While this is a concise way of describing $P^\sharp$, it offers little intuition on what the state space looks like. An alternative view is to consider the following notion of clustering. Given an index set S, a *t-partition* or *t-clustering* of S, is a set of t nonempty, disjoint subsets whose union is S. Now define $\Omega_t(n)$ to be the set of all t-partitions of $\{1, \cdots, n\}$ and $\Omega_{\leq k}(n) = \Omega_1(n) \cup \ldots \cup \Omega_k(n)$. The following lemma, whose proof appears in the appendix, establishes an alternate form of the projected Gibbs sampler and its stationary distribution.

Lemma 5.4. *The state space $\Omega_{\leq k}(n)$ is isomorphic to the set of equivalence classes induced by $\sim$ over $\{1, \ldots, k\}^n$, $\Omega^\sharp$. Furthermore, the projected Gibbs sampler specified in Figure 5.2 is the exactly the chain induced by taking the equivalence classes of the states of the collapsed Gibbs sampler. Finally, projected Gibbs sampler is reversible with respect to*

$$\pi^\flat(\mathbb{C}) \propto \frac{1}{(k - |\mathbb{C}|)!} \prod_{S \in \mathbb{C}} \frac{\Gamma(|S| + \alpha)}{\Gamma(\alpha)} q(S).$$

The $1/(k - |\mathbb{C}|)!$ term appears because $\mathbb{C}$ has $k!/(k - |\mathbb{C}|)!$ counterparts in the labeling space. The upshot of Lemma 5.4 is that $P^\sharp$ and $P^\flat$ are the same Markov chain.

5.3 Mixtures of Gaussians

We are particularly interested in mixtures of d-dimensional spherical Gaussians with known variance σ^2. A commonly used prior for this situation is the d-dimensional spherical Gaussian, due to conjugacy. Thus, we will focus on the following generative process.

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \mu_1, \ldots, \mu_k &\sim N(\mu_0, \sigma_0^2 I_d) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim N(\mu_{z_i}, \sigma^2 I_d) \end{aligned} \tag{5.3}$$

The following lemma seen, for example, in [73] establishes the conjugacy of the prior and posterior in (5.3) and gives an explicit form for the posterior.

Lemma 5.5 (Chapter 4 [73])**.** *Suppose $\mathcal{P}(\theta)$ is a family of spherical Gaussians with fixed variance σ^2 and mean θ, and our prior on θ is another spherical Gaussian with mean μ_0 and variance σ_0^2. If we observe data $y = (y_1, \ldots, y_n)$ and let $S = \{1, \ldots, n\}$, then our posterior is also a spherical Gaussian with mean μ_S and variance σ_S^2 where*

$$\mu_S = \mu_0 \cdot \frac{\sigma^2}{\sigma^2 + \sigma_0^2|S|} + \mu(S) \cdot \frac{\sigma_0^2|S|}{\sigma^2 + \sigma_0^2|S|}$$
$$\sigma_S^2 = \sigma_0^2 \cdot \frac{\sigma^2}{\sigma^2 + \sigma_0^2|S|}$$

where $\mu(S) = \frac{1}{|S|}\sum_{i\in S} y_i$ is the mean of y. Note that σ_S only depends on the cardinality of S. Further, if $\sigma_0^2 \geq \sigma^2$, the second equality immediate implies $\sigma_S^2 \in \left[\frac{\sigma^2}{|S|+1}, \frac{\sigma^2}{|S|}\right]$.

Recall that for a set of indices S, $q(S)$ is the expected probability of S under $\theta \sim \mathcal{Q}(\beta)$. In the case of Gaussians, we can work out q in closed form.

Lemma 5.6. *Let $\sigma^2, \mu_0, \sigma_0^2, Q_\beta, P_\theta, x$ be as given above. Then for any set of indices $S \subset \{1, \ldots, n\}$, we have $q(S) = L(S)R(S)$ where $L(S)$ is the probability assigned to S by the max-likelihood model,*

$$L(S) = \left(\frac{1}{2\pi\sigma^2}\right)^{|S|d/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i\in S}\|x_i - \mu(S)\|^2\right),$$

and $R(S)$ penalizes how far $\mu(S)$ is from μ_0:

$$R(S) = \left(\frac{\sigma^2}{\sigma^2 + |S|\sigma_0^2}\right)^{d/2} \exp\left(-\frac{|S|\|\mu_0 - \mu(S)\|^2}{2(\sigma^2 + |S|\sigma_0^2)}\right).$$

Lemma 5.6, whose proof appears in the appendix, also gives us a nice expression for $\Delta(\cdot,\cdot)$, which is one of the factors in the transition probabilities from Lemma 5.1.

Lemma 5.7. *Let x be as above and let $S \subset \{1, \ldots, n\}$ and $i \in \{1, \ldots, n\} \setminus S$, then*

$$\Delta(S, i) = \left(\frac{1}{2\pi(\sigma^2 + \sigma_S^2)}\right)^{d/2} \exp\left(-\frac{1}{2} \cdot \frac{\|x_i - \mu_S\|^2}{\sigma^2 + \sigma_S^2}\right).$$

And if $i \in S$, then letting $S' = S \setminus \{i\}$,

$$\Delta(S, i) = \left(\frac{1}{2\pi(\sigma^2 + \sigma_{S'}^2)}\right)^{d/2} \exp\left(-\frac{1}{2} \cdot \frac{\|x_i - \mu_{S'}\|^2}{\sigma^2 + \sigma_{S'}^2}\right).$$

Proof. Note that our function q is actually dependent on the prior $\beta = (\mu_0, \sigma_0^2)$. So for the rest of the proof, let us denote q_γ as the function q with prior γ, similarly for Q_γ. Let $\beta \circ S$ denote the posterior parameters of observing data S with prior β. That is $\beta \circ S = (\mu_S, \sigma_S^2)$ where μ_S and σ_S^2 were defined in Lemma 5.5. Then we claim that for $i \notin S$, $\Delta(S, i) = q_{\beta \circ S}(\{i\})$. To see why this is, note that by Bayes' rule

$$\frac{Q_\beta(\theta) P_\theta(S)}{\int_\Theta Q_\beta(\theta') P_{\theta'}(S) d\theta'} = Q_{\beta \circ S}(\theta).$$

Thus, we can see

$$\Delta(S, i) = \frac{q_\beta(S \cup \{i\})}{q_\beta(S)} = \frac{\int_\Theta Q_\beta(\theta) P_\theta(S) P_\theta(\{i\}) d\theta}{\int_\Theta Q_\beta(\theta') P_{\theta'}(S) d\theta'} = \int_\Theta Q_{\beta \circ S}(\theta) P_\theta(\{i\}) d\theta = q_{\beta \circ S}(\{i\}).$$

Applying Lemma 5.6 completes the proof of the first claim. To prove the second claim, we can apply the first claim to the following identity:

$$\Delta(S, i) = \frac{q(S)}{q(S \setminus \{i\})} = \frac{q(S' \cup \{i\})}{q(S')} = \Delta(S', i). \qquad \square$$

In the Bayesian setting, we typically set σ_0^2 to be large, allowing flexibility in the placement of means. To enforce this, we will require that $\sigma_0 \geq \sigma$. Additionally, μ_0 is typically set to be the origin, giving us the simplified form of μ_S:

$$\mu_S = \mu(S) \cdot \frac{\sigma_0^2 |S|}{\sigma^2 + \sigma_0^2 |S|}.$$

5.4 Lower bounds on the mixing rate

We analyze the mixing time of the projected Gibbs sampler for two cases. In the first case, the number of Gaussians is misspecified. Even though we cannot expect the Gibbs sampler to recover the correct Gaussians in this case, it still makes sense to consider the samples generated by the Markov chain and evaluate how quickly these approach the stationary distribution. The lower bound we achieve is exponential in the ratio of the intercluster distances and the variance. It is worth noting that the larger this ratio is, the more well-separated the clusters are.

The second case is the more natural case where the number of Gaussians is correctly specified. We show the mixing time of the Gibbs sampler in this case is lower bounded by the minimum of two quantities, an exponential term much like the first case and a term of the form $n^{\Omega(\alpha)}$ where α is the sparsity parameter of the Dirichlet prior.

5.4.1 Misspecified number of clusters

In our misspecified setting, we consider a sequence of points corresponding to 6 spherical clusters, $T_1, \ldots, T_6$, of n points each with diameter δr whose means are located at the vertices of a triangular prism whose edge lengths are identically r. Let S_k denote the indices of the points in cluster T_k and let our state space be $\Omega = \Omega_{\leq 3}(6n)$. Figure 5.3 displays our point configuration X_M.

The following is our main result for the misspecified setting.

Theorem 5.8. *Let $0 < \delta \leq 1/32$, $\alpha > 0$, $0 < \sigma \leq \sigma_0$, and $k = 3$. Then there is a constant $n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ s.t. for $n \geq n_0$ the mixing rate of the projected Gibbs sampler with*

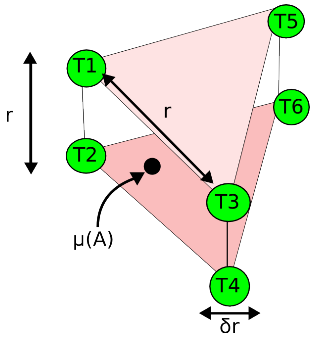


Figure 5.3. The sequence of points X_M in $\mathbb{R}^3$.

parameters α, σ, σ_0, and k over Ω is bounded below as $\tau_{mix} \geq \frac{1}{24} \cdot e^{\frac{r^2}{8\sigma^2}}$.

The full proof of Theorem 5.8 appears in the appendix, but we sketch its proof here.

The proof uses a conductance argument. Let $A = S_3 \cup \ldots \cup S_6$, and consider the singleton set V whose only element is the partition $\mathbb{C} = \{S_1, S_2, A\}$. Because of the symmetric nature of Ω, we have that $\pi(V) \leq 1/2$.

Note two properties of $\mathbb{C}$. First, the number of points in each cluster of $\mathbb{C}$ is within a constant fraction of any other cluster of $\mathbb{C}$. Second, all the points in a cluster of $\mathbb{C}$ are closer to that cluster's mean than to any other cluster's mean by a constant fraction.

To bound the conductance of V, we bound the probability that we transition out of V. This can happen in one of three ways: we can move an index in A to one of S_1 or S_2, we can move an index in S_1 or S_2 to A, or we can move an index between S_1 and S_2.

Recalling the transition probabilities from Figure 5.2 and the form of $\Delta(\cdot, \cdot)$ from Lemma 5.7, we can see the likelihood of moving a point i in a cluster S in $\mathbb{C}$ to another cluster T in $\mathbb{C}$ is roughly of the following form.

$$\begin{aligned}\Pr(\text{move } i \text{ to } T) &= \frac{(\alpha + |T|)\Delta(T, i)}{\sum_{T' \in \mathbb{C}}(\alpha + |T' \setminus \{i\}|)\Delta(T', i)} \\ &\leq \frac{(\alpha + |T|)\Delta(T, i)}{(\alpha + |S \setminus \{i\}|)\Delta(S, i)}\end{aligned}$$

$$\approx \left(\frac{\alpha + |T|}{\alpha + |S \setminus \{i\}|}\right)\left(\frac{\sigma^2 + \sigma^2_{S\setminus\{i\}}}{\sigma^2 + \sigma^2_T}\right)^{d/2}$$
$$\cdot \exp\left(\frac{\|x_i - \mu(S)\|^2}{\sigma^2} - \frac{\|x_i - \mu(T)\|^2}{\sigma^2}\right)$$

In our setup, the sizes of S and T are within a constant fraction of each other, which implies by Lemma 5.5 that the first two terms in the last line approach constants as the number of points grows. Since all the points are closer to their own cluster's mean than to any other cluster's mean by a constant fraction, the last term in the above is exponential in $-r^2/\sigma^2$. Theorem 5.8 follows by applying Theorem 4.2.

5.4.2 Correctly specified number of clusters

In our correctly specified setting, we consider a sequence of points corresponding to 3 spherical clusters, T_1, T_2, and T_3, of n points each with diameter δr whose means are located at the vertices of an equilateral triangle of edge length r and centered about the origin. Figure 5.4 displays our point configuration X_G in $\mathbb{R}^2$.

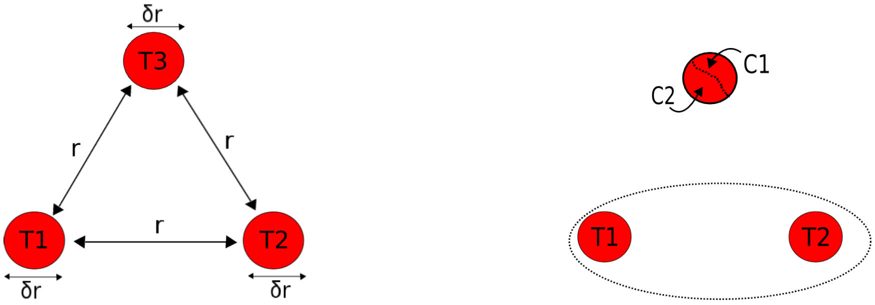


Figure 5.4. *Left*: The sequence of points X_G in $\mathbb{R}^2$. *Right*: Typical clustering in V.

Letting $\Omega = \Omega_{\leq 3}(3n)$ be our state space, we have the following result about the mixing time of P over Ω.

Theorem 5.9. *For* $\delta < \frac{1}{4}\left(\sqrt{\frac{7}{3}} - \frac{3}{2}\right)$, $\alpha \geq 1$, $0 < \sigma \leq \sigma_0$, *and* $k = 3$, *there exists* $n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ *such that* $n \geq n_0$ *implies that the mixing rate of the projected Gibbs* *sampler with parameters α, σ, σ_0, and k over Ω is bounded below as*

$$\tau_{mix} \geq \frac{1}{8}\min\left(\frac{1}{6}e^{\left(\frac{r^2}{96\sigma^2}\right)}, \frac{n^{\alpha-d/2}\left(\frac{\sigma}{\sigma_0}\right)^d \exp\left(\frac{\alpha-\alpha^2}{n}\right)}{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{r^2}{\sigma_0^2}\right)}\right).$$

To establish this result, we consider the partitions $V \subset \Omega$ such that S_1 and S_2 are clustered together and their cluster contains no indices from S_3. A typical element of V is shown in Figure 5.4. Because of the symmetric nature of Ω, we know $\pi(V) \leq 1/2$. Thus we can use the conductance of V to bound the mixing time.

Ideally, we would like to give a conductance-based argument similar to our misspecified setting. However, there is a special case to consider. V contains a special clustering where there are two clusters: $\mathbb{C} := \{S_1 \cup S_2, S_3\}$. The probability of transitioning from $\mathbb{C}$ to a clustering in V^c cannot be bounded from above in the same manner as before since we can choose a point in $S_1 \cup S_2$ and make it a singleton cluster with relatively high probability. Thus, to analyze $\Phi(V)$, we will consider V as the disjoint union of two sets $A = \{\mathbb{C}\}$ and $B = V \setminus A$. Then by the definition of conductance,

$$\Phi(V) \leq \frac{\pi(A)}{\pi(V)} + \frac{1}{\pi(V)}\sum_{x \in B, y \in V^c} \pi(x)P(x,y). \tag{5.4}$$

The following two lemmas bound each term on the right separately.

Lemma 5.10. *For $n \geq 2$ and $\alpha \geq 1$,*

$$\frac{\pi(A)}{\pi(V)} \leq \frac{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{\alpha^2-\alpha}{n} + \frac{r^2}{\sigma_0^2}\right)\sigma_0^d}{\sigma^d n^{\alpha-d/2}}.$$

Lemma 5.11. *For $\delta \leq \frac{1}{4}\left(\sqrt{\frac{7}{3}} - \frac{3}{2}\right)$, there exists an $n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ s.t. for $n \geq n_0$,*

$$\frac{1}{\pi(V)}\sum_{x \in B, y \in V^c} \pi(x)P(x,y) \leq 6\exp\left(-\frac{r^2}{96\sigma^2}\right).$$

Theorem 5.9 follows from (5.4) and Lemmas 5.10 and 5.11.

5.5 Simulations

We ran simulations of the Gibbs sampler in this setting to evaluate whether or not the bottlenecks presented above actually prevent the Gibbs sampler from finding high probability regions. For each simulation, we generated a point sequence by taking $k = 10$ draws from a d-dimensional spherical Gaussian $\mathcal{N}(0, \sigma_0^2 I_d)$ to get means $\mu_1, \ldots, \mu_{10}$. For each mean μ_i, we took $n = 50$ draws from $N(\mu_i, \sigma^2 I_d)$ with $\sigma = 0.5$.

The Gibbs sampler requires parameters $k, \alpha, \sigma^2, \sigma_0^2$ and an initial clustering. For each set of simulations, we used the same k, σ^2, and σ_0^2 that generated the point sequence over which the sampler was run. We then fixed an α and performed 10 separate runs with different initial clusterings of the points. To generate our initial configurations, we randomly chose k centers and clustered the points together that were closest to a particular center. Each run of the Gibbs sampler was done for 10^6 steps, and we plotted at each step the log of the relative probability of the current state $\mathbb{C}$.

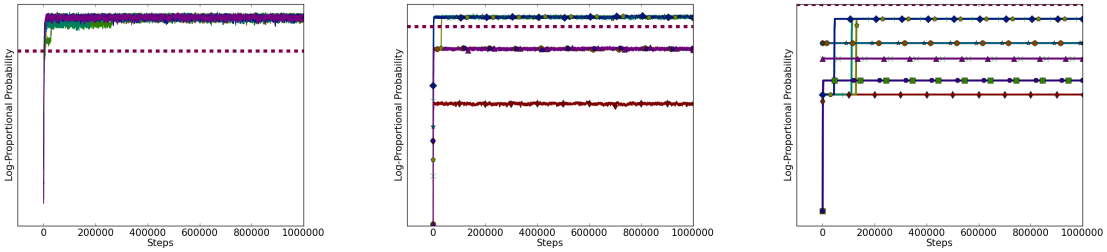
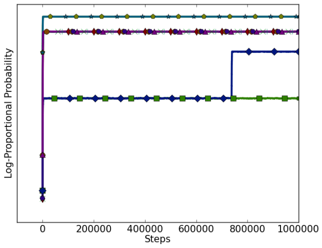


Figure 5.5. The dashed line represents the log-proportional probability of the generating clustering. *Left*: $d = 10$, $\alpha = 0.5$, $\sigma_0 = 0.5$. *Center*: $d = 3$, $\alpha = 1.5$, $\sigma_0 = 10.0$. *Right*: $d = 3$, $\alpha = 1.5$, $\sigma_0 = 1000.0$

In Figure 5.5, we can see the importance of the ratio σ_0^2/σ^2. We see that when all else is held constant, a higher value for σ_0^2/σ^2 will result in slower convergence times. Additionally, we also see that when α and σ_0^2/σ^2 are small, the Gibbs sampler will converge to a high probability state.



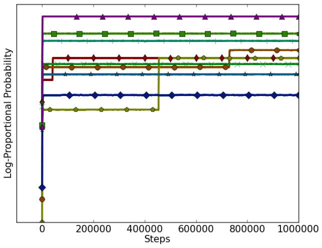
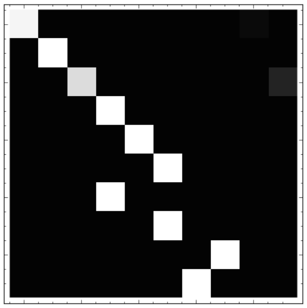


Figure 5.6. *Left*: $d = 10$, $\alpha = 1.0$, $\sigma_0 = 5.0$. *Right*: $d = 10$, $\alpha = 0.5$, $\sigma_0 = 5.0$

In Figure 5.6, we can see the importance of α. There are many more phase changes when the value of α is lower. This is possibly due to the observation in Lemma 5.10 that the relative probability mass of an empty clustering is larger when α is smaller. This makes it possible for the Gibbs sampler to create empty clusters more often and thus to make more phase transitions.

Finally, Figure 5.7 gives us an idea of what these phase transitions look like. The confusion matrices compare a clustering of the Gibbs sampler at a particular time to the generating clustering.



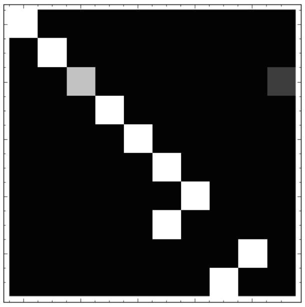


Figure 5.7. The confusion matrices of a run before and after a phase transition.

Chapter 5 contains material as it appears in “Lower bounds for the Gibbs sampler over mixtures of Gaussians.” C. Tosh and S. Dasgupta. In International Conference of Machine Learning 2014. The dissertation author was the primary investigator.

Chapter 6

Mixing rates for the alternating Gibbs sampler over Restricted Boltzmann Machines

We saw in Chapter 5 that, in the context of a Bayesian Gaussian mixture model, the Gibbs sampler can mix very slowly, even in a setting where the data looks as though it were generated by the model under consideration. In this chapter, we will investigate a variant of the Gibbs sampler, known as the alternating Gibbs sampler, in the context of a family of graphical models. We will see that in some cases, the alternating Gibbs sampler will mix rapidly, while in others, it can take an exponential amount of time to mix.

6.1 Markov Random Fields

Markov Random Fields (MRFs) are a popular class of graphical models which have found uses from image restoration [43], to modeling in statistical physics [58, 77], to pretraining deep neural networks [54, 18]. Formally, a Markov Random Field consists of an underlying graph $G = (V, E)$ and a set of random variables $X = (X_v)_{v \in V}$ indexed by the vertices V satisfying

$$P\left(X_v \,|\, X_{V \setminus \{v\}}\right) \;=\; P\left(X_v \,|\, X_{N(v)}\right)$$

where $N(v)$ is the set of vertices adjacent to v in G.

A fundamental problem in the setting of MRFs is to sample from the joint distribution $P(X)$. When the state space of X is finite and each state has positive probability, the Hammersley-Clifford theorem [50, 19] tells us that we can decompose the probability density function as

$$P(X = x) \;=\; \frac{1}{Z} \prod_{c \in \mathrm{cl}(G)} \psi_c(x_c)$$

where $\mathrm{cl}(G)$ is the set of maximal cliques of G, $\psi_c(\cdot)$ are positive functions, and Z is the normalizing constant to make the density sum to one. In general, computing Z is a hard problem [22], which makes exactly sampling from $P(X)$ challenging. The solution to this problem is to approximately sample from $P(X)$.

In the case of MRFs, the Gibbs sampler maintains a current state $(X_v = x_v)_{v \in V}$, and it takes a single step by choosing an index $v \in V$ and updating the value of X_v according to the conditional distribution $P(X_v \,|\, X_{N(v)} = x_{N(v)})$. If we can efficiently sample from these conditional distributions then each step of the Gibbs sampler is also efficient. For many MRFs of interest, this is indeed the case.

In some cases, it is possible to efficiently sample more than a single random variable at a time. Consider an MRF whose underlying graph is k-colorable, i.e. there is a partition $B_1, \ldots, B_k$ of V such that for all $i \in \{1, \ldots, k\}$ and all $u, v \in B_i$, the edge (u, v) does not appear in the graph. Then conditioning on $V \setminus B_i$, the elements of B_i are independent and the joint conditional distribution factorizes:

$$P(X_{B_i} \,|\, X_{V \setminus B_i}) \;=\; \prod_{v \in B_i} P(X_v \,|\, X_{N(v)}).$$

If we can efficiently sample from the individual conditional distributions then we can also do so for these joint conditional distributions. Moreover, we can modify the Gibbs sampler so that at each step it updates an entire block B_i, and it will still converge to the correct distribution. This Markov chain is the *alternating Gibbs sampler.*

6.1.1 Restricted Boltzmann Machines

An important special case of a Markov Random Field is the Restricted Boltzmann Machine (RBM). The underlying graph of an RBM is a fully connected bipartite graph with visible nodes $v = (v_1, \ldots, v_n)$ and hidden nodes $h = (h_1, \ldots, h_n)$. A *configuration* $x = (x(h), x(v))$ is an assignment of each node to a value in $\{0, 1\}$. The *energy* of a configuration x is

$$E(x) \ = \ -\sum_{i=1}^{n} a_i x(v_i) - \sum_{j=1}^{m} b_j x(h_j) - \sum_{i,j} x(v_i) W_{ij} x(h_j)$$

where the a_i's and b_j's are biases and the W_{ij}'s are interaction strengths or weights. This induces the Gibbs distribution over configurations: for a random configuration X, $P(X = x) = \frac{1}{Z} e^{-E(x)}$, where Z is the normalizing constant to make the distribution integrate to one. Because the underlying graph is bipartite, the conditional distribution of a visible node v_i is

$$P(X(v_i) \ = \ 1 \,|\, x(N(v_i))) \ = \ P(X(v_i) = 1 \,|\, x(h)) \ = \ \sigma\left(a_i + \sum_{j=1}^{m} W_{ij} x(h_j)\right)$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid function. Similarly, the conditional distribution of a hidden node h_j is

$$P(X(h_j) = 1 \,|\, x(v)) \ = \ \sigma\left(b_j + \sum_{i=1}^{n} W_{ij} x(v_i)\right).$$

Because these conditional distributions are easy to sample, the Gibbs sampler can be implemented efficiently.

Alternating Gibbs sampling is particularly simple in the case of RBMs. Since RBMs are built on bipartite graphs, the alternating Gibbs sampler first independently

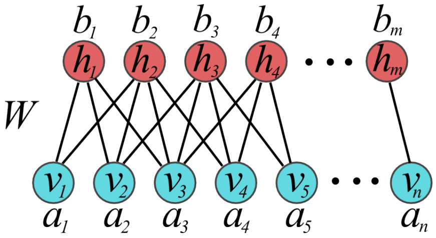


Figure 6.1. The structure of a Restricted Boltzmann Machine.

samples all of the hidden nodes conditioned on the visible nodes and then independently samples all of the visible nodes conditioned on the hidden nodes. Further, this simplicity is not restricted to RBMs themselves; it only requires that the MRF in question have an underlying graph that is bipartite.

In this chapter, we consider the mixing rates for the alternating Gibbs sampler for a wide variety of bipartite MRFs.

6.2 Preliminaries and notation

A bivariate function $d(\cdot,\cdot)$ is a *semimetric* over a space $\mathcal{X}$ if it satisfies all the properties of a metric except for the triangle inequality, i.e. non-negativity, identity iff equality, and symmetry. Any metric is trivially a semimetric. In addition, distances such as ℓ_2^2-distance are also semimetrics.

Given a matrix $A \in \mathbb{R}^{n\times m}$ and positive reals $p, q > 0$, the $L_{p,q}$-norm of A is defined as

$$\|A\|_{p,q} \;=\; \left(\sum_{j=1}^{m}\left[\sum_{i=1}^{n}|A_{ij}|^p\right]^{q/p}\right)^{1/q}.$$

We will utilize several special cases of the $L_{p,q}$-norm:

$$\|A\|_F \;=\; \|A\|_{2,2} \;=\; \sqrt{\sum_{j=1}^{m}\sum_{i=1}^{n}A_{ij}^2} \qquad \text{(Frobenius norm)}$$

$$\|A\|_1 \;=\; \|A\|_{1,\infty} \;=\; \max_{1\le j\le m}\sum_{i=1}^{n}|A_{ij}| \qquad (L_1\text{-norm})$$

$$\|A\|_{\max} \;=\; \|A\|_{\infty,\infty} \;=\; \max_{i,j} |A_{ij}| \qquad \text{(max-norm)}$$

6.3 The discrete case

Suppose that we have two vectors of nodes: visible nodes $v = (v_1, \ldots, v_n)$ and hidden nodes $h = (h_1, \ldots, h_m)$. Let $\mathcal{X}$ be some finite space, and let Ω_v denote the set of configurations x which assign to each visible node a value in $\mathcal{X}$. We can also define Ω_h to be the same except for hidden nodes and $\Omega = \Omega_v \times \Omega_h$ to be the configurations which assign to every node a value in $\mathcal{X}$.

For $x \in \Omega_h$, let $P^{(v)}(\cdot \,|\, x(h))$ denote the conditional distribution of the visible nodes given an assignment to the hidden nodes. We can symmetrically define $P^{(h)}(\cdot \,|\, x(v))$. For two configurations $x, y \in \Omega$, let $d_v(x, y)$ denote a semimetric over the assignments to the visible nodes. Similarly, let $d_h(x, y)$ denote a semimetric over the hidden nodes. Define

$$\gamma_v^{(\min)} \;=\; \min_{x \neq y} d_v(x, y) \;\text{ and }\; \gamma_v^{(\max)} \;=\; \max_{x \neq y} d_v(x, y).$$

Similarly, define $\gamma_h^{(\min)}$ and $\gamma_h^{(\max)}$ as the corresponding extremal hidden distances.

The alternating Gibbs sampler is the Markov chain $(X_t)_{t=0}^{\infty}$ taking values in Ω, which starts at some initial configuration $X_0 = x_0$, and performs the following for $t = 1, 2, \ldots$

- Draw $X_t(h) \sim P^{(h)}(\cdot \,|\, X_{t-1}(v))$
- Draw $X_t(v) \sim P^{(v)}(\cdot \,|\, X_t(h))$

We say that the distribution $P^{(v)}$ is *c-contractive* if for any assignments $x, y \in \Omega$ there exists a coupling (X, Y) of $P^{(v)}(\cdot \,|\, x(h))$ and $P^{(v)}(\cdot \,|\, y(h))$ satisfying

$$\mathbb{E}\left[d_v(X, Y)\right] \;\leq\; c\, d_h(x, y).$$

Contractivity for $P^{(h)}$ is defined symmetrically. With these notions in hand, we are ready to state our first theorem.

Theorem 6.1. *Let $c_1, c_2 \geq 0$ such that $c_1 c_2 < 1$, $P^{(v)}$ is c_1-contractive, and $P^{(h)}$ is c_2-contractive. Then the mixing rate of the Gibbs sampler is bounded as*

$$\tau(\epsilon) \ \leq \ 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{C}{\epsilon}\right)$$

where $C = \min\left(\frac{\gamma_v^{(max)}}{\gamma_v^{(min)}}, \frac{\gamma_h^{(max)}}{\gamma_h^{(min)}}, \frac{c_2 \gamma_v^{(max)}}{\gamma_h^{(min)}}\right)$.

Proof. We will prove $\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{\gamma_v^{(\max)}}{\epsilon \gamma_v^{(\min)}}\right)$. The other inequalities are left to the appendix. Our strategy is to glue together the two contractive couplings for the conditional distributions in order to make a Markovian coupling for the Gibbs sampler. Formally, if we are at time step t, then we will first sample $(X_{t+1}(h), Y_{t+1}(h))$ according to the c_1-contractive coupling of $P^{(h)}(\cdot \mid X_t(v))$ and $P^{(h)}(\cdot \mid Y_t(v))$. Then we will sample $(X_{t+1}(v), Y_{t+1}(v))$ according to the c_2-contractive coupling of $P^{(v)}(\cdot \mid X_{t+1}(h))$ and $P^{(v)}(\cdot \mid Y_{t+1}(h))$. By construction, this is a valid Markovian coupling for the alternating Gibbs sampler. For $t \geq 1$ and any initial distribution of X_0 and Y_0, we have

$$\Pr(X_t \neq Y_t) \ \leq \ \Pr(d_v(X_{t-1}, Y_{t-1}) \geq \gamma_v^{(\min)}).$$

By Markov's inequality and the law of total expectation, we have

$$\begin{aligned} \Pr(d_v(X_{t-1}, Y_{t-1}) \geq \gamma_v^{(\min)}) \ &\leq \ \frac{\mathbb{E}[d_v(X_{t-1}, Y_{t-1})]}{\gamma_v^{(\min)}} \\ &\leq \ \frac{(c_1 c_2)^{t-1} \mathbb{E}[d_v(X_0, Y_0)]}{\gamma_v^{(\min)}} \\ &\leq \ \frac{(c_1 c_2)^{t-1} \gamma_v^{(\max)}}{\gamma_v^{(\min)}} \end{aligned}$$

For $t \geq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{\gamma_v^{(\max)}}{\epsilon \gamma_v^{(\min)}}\right)$, the above is less than ϵ. Applying Lemma 4.4 completes the proof. □

6.3.1 Restricted Boltzmann Machines

Returning to the case of RBMs, recall that configurations are $\{0,1\}$-valued vectors and the conditional distributions are product distributions whose components are of the form

$$\begin{aligned} P^{(v)}_{RBM}(X(v_i) = 1 \,|\, x(h)) &= \sigma\left(a_i + \sum_{j=1}^{m} W_{ij} x(h_j)\right) \\ P^{(h)}_{RBM}(X(h_j) = 1 \,|\, x(v)) &= \sigma\left(b_j + \sum_{i=1}^{n} W_{ij} x(v_i)\right) \end{aligned}$$

where $\sigma(t) = 1/(1+\exp(-t))$ is the logistic sigmoid, and $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $W \in \mathbb{R}^{n \times m}$ are parameters of the model. We will use Hamming distance as our semimetric for both hidden and visible distances, i.e.

$$\begin{aligned} d_v(x,y) &= |\{i \,:\, x(v_i) \neq y(v_i)\}| \text{ and} \\ d_h(x,y) &= |\{j \,:\, x(h_j) \neq y(h_j)\}| \end{aligned}$$

The following lemma, which is proven in the appendix, establishes the contractivity of the RBM conditional distributions with respect to Hamming distance.

Lemma 6.2. *$P^{(v)}_{RBM}$ and $P^{(h)}_{RBM}$ are $\frac{\|W\|_1}{2}$- and $\frac{\|W^T\|_1}{2}$-contractive, respectively.*

Combining this with the simple observations that $\gamma_v^{(\min)} = \gamma_h^{(\min)} = 1$, $\gamma_v^{(\max)} = n$, and $\gamma_h^{(\max)} = m$ we have the following corollary of Theorem 6.1.

Corollary 6.3. *The mixing rate for the alternating Gibbs sampler over an RBM whose weight matrix W satisfies $\|W\|_1 \|W^T\|_1 < 4$ is upper bounded as*

$$\tau(\epsilon) \;\leq\; \frac{1}{\log(4) - \log(\|W\|_1 \|W^T\|_1)} \log\left(\frac{\min(n,m)}{\epsilon}\right).$$

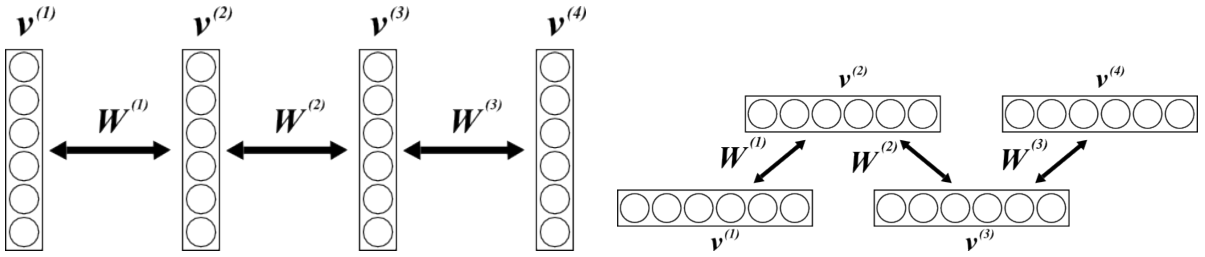


Figure 6.2. Unrolling a DBM. *Left*: the standard stacked view of a DBM. *Right*: unrolling a DBM into a bipartite graph.

6.3.2 Deep Boltzmann Machines

A natural way to generalize an RBM is to consider several stacked layers of nodes $v^{(1)}, \ldots, v^{(K)}$ of sizes $n_1, \ldots, n_K$ with interaction matrices $W^{(i)} \in \mathbb{R}^{n_i \times n_{i+1}}$ connecting them. This MRF is known as a *Deep Boltzmann Machine* (DBM) [79]. Figure 6.2 gives two visualizations of a 4-layer DBM.

As one can see from the 'unrolled' view in Figure 6.2, DBMs are also bipartite MRFs. Indeed, they are a special case of RBMs in which the visible nodes correspond to the odd layer nodes and the hidden nodes correspond to the even layer nodes and the weight matrix is given by

$$W \ := \begin{pmatrix} W^{(1)} & 0 & 0 & \\ {W^{(2)}}^T & W^{(3)} & 0 & \\ 0 & {W^{(4)}}^T & W^{(5)} & \\ 0 & 0 & & \ddots \end{pmatrix}$$

Thus the alternating Gibbs sampler can be applied to DBMs where we sample first the even layers and then the odd layers. Corollary 6.3 then immediately implies the following.

Corollary 6.4. *Let $W^{(1)}, \ldots, W^{(K)}$ be the weight matrices of a DBM and let W be defined as above. Then if $\|W\|_1 \|W^T\|_1 < 4$ the mixing rate of the alternating Gibbs sampler is* *bounded above as*

$$\tau(\epsilon) \;\leq\; \frac{1}{\log(4) - \log(\|W\|_1 \|W^T\|_1)} \log\left(\frac{\min(n,m)}{\epsilon}\right).$$

where $n = n_1 + n_3 + \cdots$ is the total number of nodes in the odd layers and $m = n_2 + n_4 + \cdots$ is the total number of nodes in the even layers.

The matrix W is far more structured in the setting of DBMs than in the setting of general RBMs, with most of its entries take the value 0. For example, if $K = 2M$, then

$$\begin{aligned}\|W\|_1 \;&=\; \max_{1 \leq k \leq M} \max_{t \in n_{2k}} \sum_{i=1}^{n_{2k-1}} |W_{it}^{(2k-1)}| + \sum_{j=1}^{n_{2k}} |W_{tj}^{(2k)}| \\ \|W^T\|_1 \;&=\; \max_{0 \leq k \leq M} \max_{t \in n_{2k+1}} \sum_{i=1}^{n_{2k}} |W_{it}^{(2k)}| + \sum_{j=1}^{n_{2k+1}} |W_{tj}^{(2k+1)}|\end{aligned}$$

where $W^{(0)}$ and $W^{(2M+1)}$ are taken to be zero matrices of the appropriate dimensions. Thus $\|W\|_1 \|W^T\|_1 < 4$ is a much less restrictive requirement in the case of DBMs than it is for general RBMs.

6.3.3 Softmax RBMs

Another way to generalize RBMs is to replace the binary logistic sigmoid units with K-ary softmax units. In this setting, the n visible units take values in $[K] = \{1, \ldots, K\}$ and the m hidden units take values in $\{0, 1\}$. Further, there are K weight matrices $W^{(1)}, \ldots, W^{(K)}$, K visible bias vectors $a^{(1)}, \ldots, a^{(K)}$, and a hidden bias vector b. Given $x \in \Omega$, the conditional distribution of a hidden node h_j is

$$P_{\mathrm{S}}^{(h)}(X(h_j) = 1 \,|\, x(v)) \;=\; \sigma\left(b_j + \sum_{i,k} W_{ij}^{(k)} \mathbf{1}[x(v_i) = k]\right).$$

For a visible node v_i, the conditional distribution is

$$P_{\mathrm{S}}^{(v)}(X(v_i) = k \,|\, x(h)) \;=\; \frac{e^{a_i^{(k)} + \sum_j x(h_j) W_{ij}^{(k)}}}{\sum_{k'=1}^{K} e^{a_i^{(k')} + \sum_j x(h_j) W_{ij}^{(k')}}}.$$

Finally, the full conditional distributions of the hidden and visible nodes are simply the product distributions. Define $W \in \mathbb{R}^{n \times m}$ as the matrix with entries

$$W_{ij} \;=\; \max_{k,k'} \left| W_{ij}^{(k)} - W_{ij}^{(k')} \right|.$$

Because there is no a priori relationship between the values in $[K]$ or in $\{0,1\}$, we will again use Hamming distance for both visible and hidden distances. The following lemma, which is proven in the appendix, establishes the contractivity of our conditional distributions.

Lemma 6.5. *$P_S^{(h)}$ and $P_S^{(v)}$ are $\frac{1}{2}\|W^T\|_1$- and $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, respectively.*

We then have the following corollary.

Corollary 6.6. *The mixing rate for the Gibbs sampler over a softmax RBM whose matrices satisfies $\binom{K}{2}\|W\|_1\|W^T\|_1 < 4$ is upper bounded as*

$$\tau(\epsilon) \;\leq\; \frac{1}{\log(4) - \log\left(\binom{K}{2}\|W\|_1\|W^T\|_1\right)} \log\left(\frac{\min(n,m)}{\epsilon}\right).$$

In the case where $K = 2$, the Softmax RBM is the original RBM in disguise. Identifying the state 1 with 0 and the state 2 with 1, taking $W = W^{(2)} - W^{(1)}$, and taking $a = a^{(2)} - a^{(1)}$ gives us the RBM conditional distributions. Thus, Corollary 6.6 is a generalization of Corollary 6.3.

6.4 The general case

We now turn our attention to a more general setting. Suppose that our vectors v and h take values in spaces Ω_v and Ω_h equipped with semimetrics $d_v(\cdot,\cdot)$ and $d_h(\cdot,\cdot)$ that do not have a minimum distance for distinct elements, i.e.

$$\inf_{x \neq y} d_v(x, y) \;=\; 0 \;=\; \inf_{x \neq y} d_h(x, y).$$

In this case, we cannot hope to apply Theorem 6.1 even if we could bound the diameter of Ω_v and Ω_h. Contractivity of the conditional distributions alone is not sufficient to guarantee rapid convergence in total variation distance. To guarantee rapid mixing, we will require another property of one of our conditional distributions. For convenience, we will use the visible conditional distribution.

Definition 6.7. *We say that $P^{(v)}$ is (ϵ, δ, M)*-gamble admissible *if for any $x, y \in \Omega$, there exists a coupling (X, Y) of $P^{(v)}(\cdot \,|\, x(h))$ and $P^{(v)}(\cdot \,|\, y(h))$ such that*

(i) $\Pr(X \neq Y \,|\, d_h\,(x, y) \leq \epsilon) \leq \delta$.

(ii) $\mathbb{E}\left[d_v(X, Y) \,|\, d_h\,(x, y) \leq \epsilon, X(v) \neq Y(v)\right] \leq M$.

(iii) $\Pr(X \neq Y \,|\, x(h) = y(h)) = 0$.

We call a coupling (X, Y) that satisfies conditions (i)-(iii) a (ϵ, δ, M)*-gamble coupling*. In contrast with the contractive couplings given in Section 6.3, a gamble coupling aims to set $X = Y$ instead of simply shrinking $d_v(X, Y)$. In particular, if $d_h\,(x, y)$ is small enough (less than ϵ), then condition (i) guarantees that $X = Y$ with probability at least $1 - \delta$. On the other hand, in the event that $X \neq Y$, condition (ii) guarantees that the expected distance between X and Y is not too large. Finally, condition (iii) guarantees that if $P^{(v)}(\cdot \,|\, x(h)) = P^{(v)}(\cdot \,|\, y(h))$, then X and Y will be the same with probability one.

The following lemma says that if both conditional distributions are contractive and one is gamble admissible, then these couplings can be interleaved in such a way to produce a Markovian coupling whose time to couple is small.

Lemma 6.8. *Let $c_1, c_2, \epsilon_0, \delta_0, M > 0$ such that $c_1 c_2 < 1$, $P^{(h)}$ is c_1-contractive, $P^{(v)}$ is c_2-contractive and $(\epsilon_0, \delta_0, M)$-gamble admissible. There exists a Markovian coupling (X_t, Y_t) such that if $\mathbb{E}[d_v(X_0, Y_0)] \leq M$, then for any $\delta > 0$, if*

$$t \;\geq\; \frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)} \log\left(\frac{2c_1M}{\delta\epsilon_0} \cdot \frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $\Pr(X_t(v) \neq Y_t(v)) \leq \delta$.

The strategy for proving Lemma 6.8 is to use our contractive coupling until $d_h(X_s, Y_s) \leq \epsilon_0$ and then apply our gamble coupling. We will succeed with probability $1 - \delta_0$, but even if we fail we are no worse off than when we started in expectation. Therefore, we can repeat this process until we achieve convergence, roughly $\frac{\log(1/\delta)}{\log(1/\delta_0)}$ times. The full proof appears in the appendix.

Unfortunately, we can not simply use Lemma 6.8 along with Lemma 4.4 to get upper bounds on the mixing rate due to the unbounded nature of our state space. That is, so long as our conditional distributions have contractivity greater than 0, for any $T \in \mathbb{N}$ and $\delta \in (0, 1)$, there may exist an initial pair of states x, y such that $\Pr(X_T \neq Y_T \mid X_0 = x, Y_0 = y) > 1 - \delta$ under the coupling (X_t, Y_t) in Lemma 6.8.

Therefore, to get bounds on the rate of convergence, we assume that the initial state of the alternating Gibbs sampler is close enough to a random state drawn from the stationary distribution in expectation. When this assumption is made, the following theorem tells us how quickly we converge to the stationary distribution.

Theorem 6.9. *Let c_1, c_2, ϵ_0, δ_0, M, $P^{(h)}$, and $P^{(v)}$ satisfy the conditions of Lemma 6.8. If X_t is the Gibbs sampler whose initial state X_0 satifies $\mathbb{E}[d_v(X_0, Y)] \leq M$ where Y is drawn independently from the stationary distribution π, then for $\delta > 0$ and any t satisfying*

$$t \;\geq\; 1 + \frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)} \log\left(\frac{2c_1M}{\delta\epsilon_0} \cdot \frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $\|X_t - \pi\|_{TV} \leq \delta$*.*

Proof. Let (X_s, Y_s) be the Markovian coupling from Lemma 6.8. Say $Y_0 \sim \pi$ independently from X_0, then $Y_0, Y_1, \ldots \sim \pi$. Further, if at time $S \geq 0$ we have $X_S(v) = Y_S(v)$, then for any time $s \geq S+1$ we have $X_s = Y_s$. Therefore for t satisfying our lower bound and for any measurable subset $A \subset \Omega$,

$$\begin{aligned}
\Pr(X_t \in A) &\geq \Pr(X_t = Y_t, Y_t \in A) \\
&\geq 1 - (\Pr(X_t \neq Y_t) + \Pr(Y_t \notin A)) \\
&\geq \Pr(Y_t \in A) - \Pr(X_{t-1}(v) \neq Y_{t-1}(v)) \\
&\geq \pi(A) - \delta.
\end{aligned}$$

Where we used Lemma 6.8 to bound $\Pr(X_{t-1}(v) \neq Y_{t-1}(v))$. Since the above holds for any measurable subset A, we can conclude $\|X_t - \pi\|_{TV} \leq \delta$. □

6.4.1 Gaussian RBMs

We now turn our attention to two special cases of continuous-valued RBMs: Gaussian-Gaussian RBMs and Gaussian-NReLU RBMs. In both cases, our configurations take values in $\mathbb{R}$.

For the Gaussian-Gaussian RBM, we have a weight matrix $W \in \mathbb{R}^{n \times m}$, bias vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, variance vectors $\sigma^2 \in \mathbb{R}^n$ and $s^2 \in \mathbb{R}^m$, and the conditional distributions are all independent normal:

$$\begin{aligned}
P_{GG}^{(v)}(X(v_i) \,|\, x(h)) &= N\left(a_i + \sum_{j=1}^{m} W_{ij}\, x(h_j),\, \sigma_i^2\right) \\
P_{GG}^{(h)}(X(h_j) \,|\, x(v)) &= N\left(b_j + \sum_{i=1}^{n} W_{ij}\, x(v_i),\, s_j^2\right)
\end{aligned}$$

For the Gaussian-NReLU RBM, the parameters W, a, and σ^2, and the conditional distribution for the visible nodes $P^{(v)}_{GN}$ are the same as in the Gaussian-Gaussian RBM. However, the hidden conditional distribution $P^{(h)}_{GN}$ has changed so that all $X(h_j)$ are independently distributed according to the noisy rectified linear distribution $\mathcal{R}(\sum_{i=1}^{n} W_{ij}\, x(v_i))$ [74], where if $Z \sim N(z, \sigma(z))$, then $\max(0, Z)$ is distributed according to $\mathcal{R}(z)$.

For both cases, the visible and hidden semimetrics that we will use will be ℓ_2^2-distance, i.e.

$$d_v(x, y) \;=\; \sum_{i=1}^{n} (x(v_i) - y(v_i))^2$$

for configurations x, y. Similarly for $d_h(x, y)$. The following lemma, whose proof appears in the appendix, establishes contractivity and gamble-admissibility for the conditional distributions we have defined.

Lemma 6.10. *The following holds.*

(a) $P^{(v)}_{GG}$*,* $P^{(h)}_{GG}$*,* $P^{(v)}_{GN}$ *are* $\|W\|_F^2$*-contractive.*

(b) $P^{(h)}_{GN}$ *is* $\frac{5}{4}\|W\|_F^2$*-contractive.*

(c) $P^{(v)}_{GG}$ *and* $P^{(v)}_{GN}$ *are* $(\epsilon_0, \delta_0, M)$*-gamble admissible for* $\epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$*,* $\delta_0 = 1/4$*, and*

$$M \;=\; 4\|\sigma\|_2^2 + \sqrt{\frac{2}{\pi}\frac{\left\|(W\sigma)^T\right\|_{2,1}}{\|(W/\sigma)^T\|_{2,1}}} + \left(\frac{\|W\|_F}{2\|(W/\sigma)^T\|_{2,1}}\right)^2$$

where W/σ *and* $W\sigma$ *denote* $n \times m$ *matrices whose entries are* W_{ij}/σ_i *and* $W_{ij}\sigma_i$*, respectively*

Lemma 6.10 and Theorem 6.9 imply the following corollary on the mixing rate for the alternating Gibbs sampler over Gaussian-Gaussian RBMs and Gaussian-NReLU RBMs.

Corollary 6.11. *Let* M *be the quantity given in Lemma 6.10. Let* X_t *denote the Gibbs sampler for the Gaussian-Gaussian RBM with stationary distribution* π_X *and* Y_t *denote the Gibbs sampler for the Gaussian-NReLU RBM with stationary distribution* π_Y*. If there* *exists $M^* > 0$ such that*

$$\max\left(\mathbb{E}_{X\sim\pi_X}\left[d_v(X_0, X)\right], \mathbb{E}_{Y\sim\pi_Y}\left[d_v(Y_0, Y)\right], M\right) \;\leq\; M^*,$$

then for $\delta > 0$ and

$$C \;=\; \frac{M^* \|(W/\sigma)^T\|_{2,1}^2 \|W\|_F^2 \log(\frac{2}{\delta})}{\delta \log(4)},$$

(a) if $\|W\|_F \leq 1$ and

$$t \;\geq\; 1 + \frac{\log(2/\delta)\log\left(8C\right)}{\log\left(\frac{1}{\|W\|_F^4}\right)\log(4)},$$

then $\|X_t - \pi_X\|_{TV} \leq \delta$, and

(b) if $\|W\|_F^4 \leq 4/5$ and

$$t \;\geq\; 1 + \frac{\log(2/\delta)\log\left(10C\right)}{\log\left(\frac{4}{5\|W\|_F^4}\right)\log(4)},$$

then $\|Y_t - \pi_Y\|_{TV} \leq \delta$.

6.5 Lower bounds

We now turn our attention towards providing lower bounds for the mixing rate of the alternating Gibbs sampler. Our approach utilizes the method of conductance, the same technique that was applied in Chapter 5 to provide lower bounds on the Gibbs sampler for Bayesian Gaussian mixture models.

Our first lower bound is for the case of RBMs.

Theorem 6.12. *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there is a weight matrix $W \in \mathbb{R}^{n\times m}$ satisfying*

$$\|W\|_{max} \;\leq\; \frac{2}{\min(n,m)} \ln\left(4T(n+m)\right)$$

such that the Gibbs sampler over the RBM with zero bias and weight matrix W has mixing rate bounded as $\tau_{mix} \geq T$.

The proof of Theorem 6.12 appears in the appendix, but we present the main idea here. We construct a weight matrix W such that the energy function associated with W has two antipodal global minima. Because there are two minima, the singleton set consisting of one minima has probability mass less than 1/2 under the Gibbs distribution. Escaping from one of these minima is a very unlikely event, implying that the conductance is small and therefore the mixing rate is large.

Our second lower bound is for the case of Gaussian-Gaussian RBMs. The state space of a Gaussian-Gaussian RBM is unbounded, but any implementation of the Gibbs sampler is necessarily in a bounded state space. Therefore, lower bounds that exploit the unbounded nature of the state space may not be particularly meaningful. To compensate for this, we work with a restricted version of the alternating Gibbs sampler. Given $B > 0$, consider the following *B-thresholded* alternating Gibbs sampler $(Y_t)_{t=0}^{\infty}$. At time step t, it performs the following.

1. For each hidden node h_j, draw $X_t(h_j) \sim N(b_j + \sum_{i=1}^{n} W_{ij}\, Y_{t-1}(v_i),\, s_j^2)$. Set $Y_t(h_j)$ to be the closest point in $[-B, B]$ to $X_t(h_j)$.

2. For each hidden node h_j, draw $X_t(v_i) \sim N(a_i + \sum_{j=1}^{m} W_{ij}\, Y_t(h_j),\, \sigma_i^2)$. Set $Y_t(v_i)$ to be the closest point in $[-B, B]$ to $X_t(v_i)$.

The following theorem, whose proof appears in the appendix, gives a lower bound on the mixing rate for this restricted Markov chain.

Theorem 6.13. *Let $T, B > 0$ and $n, m \in \mathbb{N}$ be even positive integers. Then there exists weight matrix $W \in \mathbb{R}^{n \times m}$ s.t.*

$$\|W\|_{max} \;\leq\; \frac{1}{\min(n,m)}\left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$$

such that the B-truncated chain of the Gibbs sampler for the Gaussian-Gaussian RBM with no biases and unit variances mixes in time $\tau_{mix} \geq T$.

In the case where $n = m$, the restriction on W translates to a $1 + \frac{1}{B}\sqrt{8\log(4Tn)}$ upper bound on the Frobenius norm of W. This implies that for any $\epsilon, T > 0$, there exists a $B > 0$ and a weight matrix W such that $\|W\|_F \leq 1 + \epsilon$, but the alternating Gibbs sampler mixes in time bounded below by T. In this sense, the condition on the Frobenius norm of W given in Corollary 6.11(a) is tight for establishing finite convergence rates on the alternating Gibbs sampler over Gaussian-Gaussian RBMs.

6.6 Complexity of RBMs

The results in the previous sections give conditions under which a particular algorithm, the alternating Gibbs sampler, can efficiently sample from the Gibbs distributions of RBMs and several of its variants. It is natural to ask how much better can we hope to do with either a better analysis of the Gibbs sampler or a different algorithm altogether.

The complexity of approximately sampling from a distribution is often closely tied to the complexity of approximately computing its normalizing constant or partition function [62, 69]. Therefore, to help understand the complexity of sampling from the Gibbs distribution over RBMs, we will focus on the complexity of computing approximate solutions to the following problem.

Restricted Boltzmann Machine partition: #RBM
Input: Parameters $W \in \mathbb{R}^{n\times m}$, $a \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$.
Output: The partition function

$$Z = \sum_{x\,:\,(v,h)\to\{0,1\}^{n+m}} e^{a^T x(v) + b^T x(h) + x(v)^T W x(h)}.$$

In the complexity literature, there are three well-documented categories that an approximate counting problem can be placed in. The first category consists of problems for which we have an efficient algorithm to approximately count or compute a partition function. The second category consists of problems for which an efficient approximate counting algorithm would imply the equivalence of two complexity classes widely viewed to be distinct, such as P and NP. Finally, problems in the third category do not belong to either of the above categories but often are placed in well-defined classes of possibly intermediate computational complexity. As we shall see, #RBM exhibits flavors of all three of these categories.

When all weights are positive and all biases are consistent, there is an efficient algorithm to approximate #RBM [61]. Moreover, we can combine our results from Section 6.3 with annealing techniques [83] to get an efficient algorithm for the general case when $\|W\|_1 \|W^T\|_1 < 4$. Putting this all together, we have the following result which place certain instances of #RBM into the first category.

Theorem 6.14. ([61], this chapter, [83]) *#RBM admits an efficient algorithm in both of the following cases.*

(i) $\forall\, (i, j) \in [n] \times [m]$, $W_{ij} \geq 0$ *and* $\mathit{sign}(a_i) = \mathit{sign}(b_j)$.

(ii) $\|W\|_1 \|W^T\|_1 < 4$.

On the other hand, [69] showed that when the max-norm of the weight matrix grows quickly enough, #RBM falls into the second category.

Theorem 6.15 (Long and Servedio [69])**.** *There is a universal constant* $\alpha > 0$ *such that if* $P \neq NP$*, then there is no polynomial-time algorithm such that given an* $n \times n$ *matrix* W *such that* $\|W\|_{max} \leq \psi(n) = \omega(n)^1$*, the algorithm approximates #RBM with weight matrix* W *and no bias to within a multiplicative factor of* $e^{\alpha\psi(n)}$.

Finally, [45] showed that when the weights are constrained to be positive but the biases may be arbitrary, #RBM falls into the third category. Formally, they showed that

[1] Two functions $f(n), g(n)$ satisfy the relationship $g(n) = \omega(f(n))$ if $\lim_{n\to\infty} \frac{g(n)}{f(n)} = \infty$.

it belongs to a class of problems introduced by [39] that are approximation-preserving interreducible[2] with the problem of counting independent sets in bipartite graphs (#BIS).

Theorem 6.16. ([45]) *#BIS* $\equiv_{AP}$ *#RBM when* $W_{ij} \geq 0$ *and* $\|W\|_1 \|W^T\|_1 = \Omega(n^2)$.

Theorems 6.15 and 6.16 both imply that for large values of $\|W\|_1 \|W^T\|_1$, it seems unlikely that we will be able to sample from the Gibbs distribution over RBMs, even when all the weights are constrained to be positive. On the other hand, Theorem 6.14 gives hope that there are cases when we can succeed. However, there are large gaps in the cases that we know can be efficiently solved and those in which we believe that they cannot. Closing these gaps remains an interesting direction for future research.

6.7 Related work

There has been some recent work on proving mixing rates for the Gibbs sampler on a wide range of models. Notably, [68, 37, 47] gave upper bounds for the mixing rate for the single-site update Gibbs sampler over a wide class of models which include certain discrete-valued MRFs.

[37, 47] both introduced quantities for the models that they consider for which the mixing rate of the Gibbs sampler is polynomial in the size of the model and exponential in these special quantities. More closely related to this chapter, [68, 47] also showed that if the model meets a certain 'bounded influence' criterion, then the single-site update Gibbs sampler mixes in time $O(n \log n)$.

There has also been some recent work on single-site Gibbs sampling in general state spaces. Notably, [88] gave general convergence rates for the single-site update Gibbs sampler on general state spaces.

In this chapter, we also gave general convergence results in both discrete and continuous spaces, but for the alternating Gibbs sampler as opposed to the single-site

[2] For a precise definition of approximation-preserving reducibility, see [39].

update Gibbs sampler. We applied these results to a variety of models closely related to the standard RBM, such as the Gaussian-NReLU RBM, for which mixing rate bounds were previously unknown.

Chapter 6 contains material as it appears in “Mixing Rates for the alternating Gibbs sampler over Restricted Boltzmann Machines and friends.” C. Tosh. In International Conference of Machine Learning 2016. The dissertation author was the primary investigator.

Part III

Interactive learning

Chapter 7

Diameter-based active learning

In the remainder of this thesis, we turn our focus to learning situations in which a learning algorithm is allowed to solicit interaction from a user. In this chapter, we examine a classical interactive learning setting, known as active learning, and present an algorithm with useful theoretical guarantees.

7.1 Active learning

In many situations where a classifier is to be learned, it is easy to collect unlabeled data but costly to obtain labels. This has motivated the *pool-based active learning* model, in which a learner has access to a collection of unlabeled data points and is allowed to ask for individual labels in an adaptive manner. The hope is that choosing these queries intelligently will rapidly yield a low-error classifier, much more quickly than with random querying. A central focus of active learning is developing efficient querying strategies and understanding their label complexity.

Over the past decade or two, there has been substantial progress in developing such rigorously-justified active learning schemes for general concept classes. For the most part, these schemes can be described as *mellow*: rather than focusing upon maximally informative points, they query any point whose label cannot reasonably be inferred from the information received so far. It is of interest to develop more aggressive strategies with better label complexity.

An exception to this general trend is the aggressive strategy of [31], whose label complexity is known to be optimal in its dependence on a key parameter called the *splitting index*. However, this strategy has been primarily of theoretical interest because it is difficult to implement algorithmically. In this paper, we introduce a variant of the methodology that yields efficient algorithms. We show that it admits roughly the same label complexity bounds as well as having promising experimental performance.

As with the original splitting index result, we operate in the *realizable* setting, where data can be perfectly classified by some function h^* in the hypothesis class $\mathcal{H}$. At any given time during the active learning process, the remaining candidates—that is, the elements of $\mathcal{H}$ consistent with the data so far—are called the *version space*. The goal of aggressive active learners is typically to pick queries that are likely to shrink this version space rapidly. But what is the right notion of size? Dasgupta [31] pointed out that the *diameter* of the version space is what matters, where the distance between two classifiers is taken to be the fraction of points on which they make different predictions. Unfortunately, the diameter is a difficult measure to work with because it cannot, in general, be decreased at a steady rate. Thus the earlier work used a procedure that has quantifiable label complexity but is not conducive to implementation.

We take a fresh perspective on this earlier result. We start by suggesting an alternative, but closely related, notion of the size of a version space: the *average* pairwise distance between hypotheses in the version space, with respect to some underlying probability distribution π on $\mathcal{H}$. This distribution π can be arbitrary—that is, there is no requirement that the target h^* is chosen from it—but should be chosen so that it is easy to sample from. When $\mathcal{H}$ consists of linear separators, for instance, a good choice would be a log-concave density, such as a Gaussian.

At any given time, the next query x is chosen roughly as follows:

- Sample a collection of classifiers $h_1, h_2, \ldots, h_m$ from π restricted to the current version space V.

- Compute the distances between them; this can be done using just the unlabeled points.

- Any candidate query x partitions the classifiers $\{h_i\}$ into two groups: those that assign it a + label (call these V_x^+) and those that assign it a − label (call these V_x^-). Estimate the average-diameter after labeling x by the sum of the distances between classifiers h_i within V_x^+, or those within V_x^-, whichever is larger.

- Out of the pool of unlabeled data, pick the x for which this diameter-estimate is smallest.

This is repeated until the version space has small enough average diameter that a random sample from it is very likely to have error less than a user-specified threshold ϵ. We show how all these steps can be achieved efficiently, as long as there is a sampler for π.

Dasgupta [31] pointed out that the label complexity of active learning depends on the underlying distribution, the amount of unlabeled data (since more data means greater potential for highly-informative points), and also the target classifier h^*. That paper identifies a parameter called the *splitting index* ρ that captures the relevant geometry, and gives upper bounds on label complexity that are proportional to $1/\rho$, as well as showing that this dependence is inevitable. For our modified notion of diameter, a different *averaged* splitting index is needed. However, we show that it can be bounded by the original splitting index, with an extra multiplicative factor of $\log(1/\epsilon)$; thus all previously-obtained label complexity results translate immediately for our new algorithm.

7.2 Related work

The theory of active learning has developed along several fronts.

One of these is *nonparametric* active learning, where the learner starts with a pool of unlabeled points, adaptively queries a few of them, and then fills in the remaining labels. The goal is to do this with as few errors as possible. (In particular, the learner does not return a classifier from some predefined parametrized class.) One scheme begins by building a neighborhood graph on the unlabeled data, and propagating queried labels along the edges of this graph [90, 26, 29]. Another starts with a hierarchical clustering of the data and moves down the tree, sampling at random until it finds clusters that are relatively pure in their labels [33]. The label complexity of such methods have typically be given in terms of smoothness properties of the underlying data distribution [25, 64].

Another line of work has focused on active learning of linear separators, by querying points close to the current guess at the decision boundary [15, 35, 16]. Such algorithms are close in spirit to those used in practice, but their analysis to date has required fairly strong assumptions to the effect that the underlying distribution on the unlabeled points is logconcave. Interestingly, regret guarantees for online algorithms of this sort can be shown under far weaker conditions [27].

The third category of results, to which this chapter belongs, considers active learning strategies for general concept classes $\mathcal{H}$. Some of these schemes [28, 34, 20, 13, 89] are fairly mellow in the sense described earlier, using generalization bounds to gauge which labels can be inferred from those obtained so far. The label complexity of these methods can be bounded in terms of a quantity known as the disagreement coefficient [51]. In the realizable case, the canonical such algorithm is that of [28], henceforth referred to as CAL. Other methods use a prior distribution π over the hypothesis class, sometimes assuming that the target classifier is a random draw from this prior. These methods typically aim to shrink the mass of the version space under π, either greedily and explicitly [30, 48, 46] or implicitly [41]. Perhaps the most widely-used of these methods is the latter, query-by-committee, henceforth QBC. As mentioned earlier, shrinking π-mass is not an optimal strategy if low misclassification error is the ultimate goal. In particular, what matters is not the prior mass of the remaining version space, but rather how *different* these candidate classifiers are from each other. This motivates using the diameter of the version space as a yardstick, which was first proposed in [31] and is taken up again here.

7.3 Preliminaries

Consider a binary hypothesis class $\mathcal{H}$, a data space $\mathcal{X}$, and a distribution $\mathcal{D}$ over $\mathcal{X}$. For mathematical convenience, we will restrict ourselves to finite hypothesis classes. (We can do this without loss of generality when $\mathcal{H}$ has finite VC dimension, since we only use the predictions of hypotheses on a pool of unlabeled points; however, we do not spell out the details of this reduction here.) The *hypothesis distance* induced by $\mathcal{D}$ over $\mathcal{H}$ is the pseudometric

$$d(h, h') \;:=\; \Pr_{x \sim \mathcal{D}}(h(x) \neq h'(x)).$$

Given a point $x \in \mathcal{X}$ and a subset $V \subset \mathcal{H}$, denote $V_x^+ = \{h \in V : h(x) = 1\}$ and $V_x^- = V \setminus V_x^+$. Given a sequence of data points $x_1, \ldots, x_n$ and a target hypothesis h^*, the induced *version space* is the set of hypotheses that are consistent with the target hypotheses on the sequence, i.e.

$$\{h \in \mathcal{H} : h(x_i) = h^*(x_i) \text{ for all } i = 1, \ldots, n\}.$$

7.3.1 Diameter and the Splitting Index

The *diameter* of a set of hypotheses $V \subset \mathcal{H}$ is the maximal distance between any two hypotheses in V, i.e.

$$\text{diam}(V) := \max_{h, h' \in V} d(h, h').$$

Without any prior information, any hypothesis in the version space could be the target. Thus the worst case error of any hypothesis in the version space is the diameter of the version space. The splitting index roughly characterizes the number of queries required for an active learning algorithm to reduce the diameter of the version space below ϵ.

While reducing the diameter of a version space $V \subset \mathcal{H}$, we will sometimes identify pairs of hypotheses $h, h' \in V$ that are far apart and therefore need to be separated. We will refer to $\{h, h'\}$ as an *edge*. Given a set of edges $E = \{\{h_1, h'_1\}, \ldots, \{h_n, h'_n\}\} \subset \binom{\mathcal{H}}{2}$, we say a data point x ρ-splits E if querying x separates at least a ρ fraction of the pairs, that is, if

$$\max\left\{|E_x^+|, |E_x^-|\right\} \;\leq\; (1-\rho)|E|$$

where $E_x^+ = E \cap \binom{\mathcal{H}_x^+}{2}$ and similarly for E_x^-. When attempting to get accuracy $\epsilon > 0$, we need to only eliminate edge of length greater than ϵ. Define

$$E_\epsilon \;=\; \{\{h, h'\} \in E \,:\, d(h, h') > \epsilon\}.$$

The *splitting index* of a set $V \subset \mathcal{H}$ is a tuple (ρ, ϵ, τ) such that $\Pr_{x\sim\mathcal{D}}(x\, \rho\text{-splits } E_\epsilon) \;\geq\; \tau$ for all finite edge-sets $E \subset \binom{V}{2}$.

The following theorem, due to Dasgupta [31], bounds the sample complexity of active learning in terms of the splitting index. The $\tilde{O}$ notation hides polylogarithmic factors in d, ρ, τ, $\log 1/\epsilon$, and the failure probability δ.

Theorem 7.1 (Dasgupta 2005)**.** *Suppose $\mathcal{H}$ is a hypothesis class with splitting index (ρ, ϵ, τ). Then to learn a hypothesis with error ϵ,*

(a) any active learning algorithm with $\leq 1/\tau$ unlabeled samples must request at least $1/\rho$ labels, and

(b) there is an active learning algorithm that draws $\tilde{O}(d/(\rho\tau)\log^2(1/\epsilon))$ unlabeled data points and requests $\tilde{O}((d/\rho)\log^2(1/\epsilon))$ labels when $\mathcal{H}$ has VC-dimension d.

Unfortunately, the only known algorithm satisfying (b) above is intractable for all but the simplest hypothesis classes: it constructs an ϵ-covering of the hypothesis space and queries points which whittle away at the diameter of this covering. To overcome this intractability, we consider a slightly more benign setting in which we have a samplable prior distribution π over our hypothesis space $\mathcal{H}$.

7.3.2 An Average Notion of Diameter

With a prior distribution, it makes sense to shift away from the worst-case to the average-case. We define the *average diameter* of a subset $V \subset \mathcal{H}$ as the expected distance between two hypotheses in V randomly drawn from π, i.e.

$$\Phi(V) := \mathbb{E}_{h,h' \sim \pi|_V}[d(h,h')]$$

where $\pi|_V$ is the conditional distribution induced by restricting π to V, that is, $\pi|_V(h) = \pi(h)/\pi(V)$ for $h \in V$.

Intuitively, a version space with very small average diameter ought to put high weight on hypotheses that are close to the true hypothesis. Indeed, given a version space V with $h^* \in V$, the following lemma shows that if $\Phi(V)$ is small enough, then a low error hypothesis can be found by two popular heuristics: random sampling and MAP estimation.

Lemma 7.2. *Suppose $V \subset \mathcal{H}$ contains h^*. Pick $\epsilon > 0$.*

(a) (Random sampling) If $\Phi(V) \leq \epsilon\, \pi|_V(h^*)$ then $\mathbb{E}_{h \sim \pi|_V}[d(h^*,h)] \leq \epsilon$.

(b) (MAP estimation) Write $p_{map} = \max_{h \in V} \pi|_V(h)$. Pick $0 < \alpha < p_{map}$. If

$$\Phi(V) \;\leq\; 2\epsilon \left(\min\{\pi|_V(h^*), p_{map} - \alpha\}\right)^2,$$

then $d(h^*,h) \leq \epsilon$ for any h with $\pi|_V(h) \geq p_{map} - \alpha$.

Proof. Part (a) follows from

$$\Phi(V) \;=\; \mathbb{E}_{h,h' \sim \pi|_V}[d(h,h')] \;\geq\; \pi|_V(h^*)\mathbb{E}_{h \sim \pi|_V}[d(h^*,h)].$$

For (b), take $\delta = \min(\pi|_V(h^*), p_{map} - \alpha)$ and define $V_{\pi,\delta} = \{h \in V \,:\, \pi|_V(h) \geq \delta\}$. Note that $V_{\pi,\delta}$ contains h^* as well as any $h \in V$ with $\pi|_V(h) \geq p_{map} - \alpha$.

We claim $\mathrm{diam}(V_{\pi,\delta})$ is at most ϵ. Suppose not. Then there exist $h_1, h_2 \in V_{\pi,\delta}$ satisfying $d(h_1,h_2) > \epsilon$, implying

$$\Phi(V) \;=\; \mathbb{E}_{h,h' \sim \pi|_V}[d(h,h')] \;\geq\; 2 \cdot \pi|_V(h_1) \cdot \pi|_V(h_2) \cdot d(h_1,h_2) \;>\; 2\delta^2\epsilon.$$

But this contradicts our assumption on $\Phi(V)$. Since both $h, h^* \in V_{\pi,\delta}$, we have (b). □

7.3.3 An Average Notion of Splitting

We now turn to defining an average notion of splitting. A data point x *ρ-average splits V* if

$$\max\left\{\frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+), \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-)\right\} \;\leq\; (1-\rho)\Phi(V).$$

And we say a set $S \subset \mathcal{H}$ has *average splitting index* (ρ, ϵ, τ) if for any subset $V \subset S$ such that $\Phi(V) > \epsilon$,

$$\mathrm{Pr}_{x \sim \mathcal{D}}\left(x \;\rho\text{-average splits } V\right) \;\geq\; \tau.$$

Intuitively, average splitting refers to the ability to significantly decrease the potential function

$$\pi(V)^2\Phi(V) \;=\; \mathbb{E}_{h,h' \sim \pi}[\mathbf{1}(h, h' \in V)\, d(h,h')]$$

with a single query.

While this potential function may seem strange at first glance, it is closely related to the original splitting index. The following lemma, whose proof is deferred to Section 7.5, shows the splitting index bounds the average splitting index for any hypothesis class.

Lemma 7.3. *Let π be a probability measure over a hypothesis class $\mathcal{H}$. If $\mathcal{H}$ has splitting index (ρ, ϵ, τ), then it has average splitting index $(\frac{\rho}{4\lceil\log(1/\epsilon)\rceil}, 2\epsilon, \tau)$.*

Dasgupta [31] derived the splitting indices for several hypothesis classes, including intervals and homogeneous linear separators. Lemma 7.3 implies average splitting indices within a $\log(1/\epsilon)$ factor in these settings.

Moreover, given access to samples from $\pi|_V$, we can easily estimate the quantities appearing in the definition of average splitting. For an edge sequence $E = (\{h_1, h'_1\}, \ldots, \{h_n, h'_n\})$, define

$$\psi(E) := \sum_{i=1}^{n} d(h_i, h'_i).$$

When h_i, h'_i are i.i.d. draws from $\pi|_V$ for all $i = 1, \ldots, n$, which we denote $E \sim (\pi|_V)^{2\times n}$, the random variables $\psi(E)$, $\psi(E_x^-)$, and $\psi(E_x^+)$ are unbiased estimators of the quantities appearing in the definition of average splitting.

Lemma 7.4. *Given $E \sim (\pi|_V)^{2\times n}$, we have*

- $\mathbb{E}\left[\frac{1}{n}\psi(E)\right] = \Phi(V)$ *and*
- $\mathbb{E}\left[\frac{1}{n}\psi(E_x^+)\right] = \frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+)$ *for any $x \in \mathcal{X}$. Similarly for E_x^- and V_x^-.*

Proof. From definitions and linearity of expectations, it is easy to observe $\mathbb{E}[\psi(E)] = n\,\Phi(V)$. By the independence of h_i, h'_i, we additionally have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\psi(E_x^+)\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{\{h_i,h'_i\}\in E_x^+} d(h_i, h'_i)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{\{h_i,h'_i\}\in E} \mathbf{1}[h_i \in V_x^+]\,\mathbf{1}[h'_i \in V_x^+]\,d(h_i, h'_i)\right]\end{aligned}$$

$$= \frac{1}{n} \sum_{\{h_i, h_i'\} \in E} \left(\frac{\pi(V_x^+)}{\pi(V)} \right)^2 \mathbb{E}\left[d(h_i, h_i') \,|\, h_i, h_i' \in V_x^+ \right]$$

$$= \left(\frac{\pi(V_x^+)}{\pi(V)} \right)^2 \Phi(V_x^+). \qquad \square$$

While it is tempting to define average splitting in terms of the average diameter as

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \;\leq\; (1-\rho)\Phi(V).$$

However, this definition does not satisfy a nice relationship with the splitting index. Indeed, there exist hypothesis classes V for which there are many points which 1/4-split E for any $E \subset \binom{V}{2}$ but for which every $x \in \mathcal{X}$ satisfies

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \;\approx\; \Phi(V).$$

This observation is formally proven in the appendix.

7.4 An Average Splitting Index Algorithm

Suppose we are given a version space V with average splitting index (ρ, ϵ, τ). If we draw $\tilde{O}(1/\tau)$ points from the data distribution then, with high probability, one of these will ρ-average split V. Querying that point will result in a version space V' with significantly smaller potential $\pi(V')^2\Phi(V')$.

If we knew the value ρ a priori, then Lemma 7.4 combined with standard concentration bounds [55, 3] would give us a relatively straightforward procedure to find a good query point:

- Draw $E' \sim (\pi|_V)^{2\times M}$ and compute the empirical estimate $\widehat{\Phi}(V) = \frac{1}{M}\psi(E')$.
- Draw $E \sim (\pi|_V)^{2\times N}$ for N depending on ρ and $\widehat{\Phi}$.

```
Version space $V$, prior $\pi$, data $\mathbf{x} = (x_1, \ldots, x_m)$
Set $\widehat{\rho}_1 = 1/2$
For $t = 1, 2, \ldots$:
    • Draw $E' \sim (\pi|_V)^{2 \times m_t}$ and compute $\widehat{\Phi}_t = \frac{1}{m_t}\psi(E')$
    • Draw $E \sim (\pi|_V)^{2 \times n_t}$
    • If $\exists\, x_i$ s.t. $\frac{1}{n_t} \max\left\{\psi(E^+_{x_i}), \psi(E^-_{x_i})\right\} \leq (1 - \widehat{\rho}_t)\widehat{\Phi}_t$, then halt and return $x_i$
    • Otherwise, let $\widehat{\rho}_{t+1} = \widehat{\rho}_t / 2$
```

Figure 7.1. Select algorithm.

- For suitable M and N, it will be the case that with high probability, for some x,

$$\frac{1}{N} \max\left\{\psi(E^+_x), \psi(E^-_x)\right\} \;\approx\; (1 - \rho)\widehat{\Phi}.$$

Querying that point will decrease the potential.

However, we typically would not know the average splitting index ahead of time. Moreover, it is possible that the average splitting index may change from one version space to the next. In the next section, we describe a query selection procedure that adapts to the splittability of the current version space.

7.4.1 Finding a Good Query Point

SELECT, displayed in Figure 7.1, is our query selection procedure. It takes as input a sequence of data points $x_1, \ldots, x_m$, at least one of which ρ-average splits the current version space, and with high probability finds a data point that $\rho/8$-average splits the version space.

SELECT proceeds by positing an optimistic estimate of ρ, which we denote $\widehat{\rho}_t$, and successively halving it until we are confident that we have found a point that $\widehat{\rho}_t$-average splits the version space. In order for this algorithm to succeed, we need to choose n_t and m_t such that with high probability (1) $\widehat{\Phi}_t$ is an accurate estimate of $\Phi(V)$ and (2) our halting condition will be true if $\widehat{\rho}_t$ is within a constant factor of ρ and false otherwise. The following lemma, whose proof is in the appendix, provides such choices for n_t and m_t.

Lemma 7.5. *Let $\rho, \epsilon, \delta_0 > 0$ be given. Suppose that version space V satisfies $\Phi(V) > \epsilon$. In* SELECT*, fix a round t and data point $x \in \mathcal{X}$ that exactly ρ-average splits V (that is, $\max\{\pi|_V(V_x^+)^2\Phi(V_x^+),\ \pi|_V(V_x^-)^2\Phi(V_x^-)\} = (1-\rho)\Phi(V)$). If*

$$m_t \geq \frac{48}{\widehat{\rho}_t^2 \epsilon} \log \frac{4}{\delta_0} \qquad \text{and} \qquad n_t \geq \max\left\{ \frac{32}{\widehat{\rho}_t^2 \widehat{\Phi}_t}, \frac{40}{\widehat{\Phi}_t^2} \right\} \log \frac{4}{\delta_0}$$

then with probability $1 - \delta_0$,

(a) $\widehat{\Phi}_t \geq (1 - \widehat{\rho}_t/4)\Phi(V)$;

(b) if $\rho \leq \widehat{\rho}_t/2$, then $\frac{1}{n_t} \max\{\psi(E_x^+), \psi(E_x^-)\} > (1 - \widehat{\rho}_t)\widehat{\Phi}_t$; and

(c) if $\rho \geq 2\widehat{\rho}_t$, then $\frac{1}{n_t} \max\{\psi(E_x^+), \psi(E_x^-)\} \leq (1 - \widehat{\rho}_t)\widehat{\Phi}_t$.

Given the above lemma, we can establish a bound on the number of rounds and the total number of hypotheses SELECT needs to find a data point that $\rho/8$-average splits the version space.

Theorem 7.6. *Suppose that* SELECT *is called with a version space V with $\Phi(V) \geq \epsilon$ and a collection of points $x_1, \ldots, x_m$ such that at least one of x_i ρ-average splits V. If $\delta_0 \leq \delta/(2m(2 + \log(1/\rho)))$, then with probability at least $1 - \delta$,* SELECT *returns a point x_i that $(\rho/8)$-average splits V, finishing in less than $\lceil \log(1/\rho) \rceil + 1$ rounds and sampling $O\left(\left(\frac{1}{\epsilon\rho^2} + \frac{\log(1/\rho)}{\Phi(V)^2}\right) \log \frac{1}{\delta_0}\right)$ hypotheses in total.*

Before we prove Theorem 7.6, two observations are in order.

First, it is possible to modify SELECT to find a point x_i that $(c\rho)$-average splits V for any constant $c < 1$ while only having to draw $O(1)$ more hypotheses in total. By halving $\widehat{\rho}_t$ at each step, we immediately give up a factor of two in our approximation. This can be made smaller by taking narrower steps. Additionally, with a constant factor increase in m_t and n_t, the approximation ratios in Lemma 7.5 can be set to any constant.

Second, though it appears that SELECT requires us to know ρ in order to calculate δ_0, a crude lower bound on ρ suffices. Such a bound can often be found in terms of ϵ. This is because any version space is $(\epsilon/2, \epsilon, \epsilon/2)$-splittable [31, Lemma 1]. By Lemma 7.3, so long as τ is less than $\epsilon/4$, we can substitute $\frac{\epsilon}{8\lceil\log(2/\epsilon)\rceil}$ for ρ in when we compute δ_0.

Proof of Theorem 7.6. Let $T := \lceil\log(1/\rho)\rceil + 1$. By Lemma 7.5, we know that for rounds $t = 1, \ldots, T$, we don't return any point which does worse than $\widehat{\rho}_t/2$-average splits V with probability $1 - \delta/2$. Moreover, in the T-th round, it will be the case that $\rho/4 \leq \widehat{\rho}_T \leq \rho/2$, and therefore, with probability $1 - \delta/2$, we will select a point which does no worse than $\widehat{\rho}_T/2$-average split V, which in turn does no worse than $\rho/8$-average split V.

Note that we draw $m_t + n_t$ hypotheses at each round. By Lemma 7.5, for each round $\widehat{\Phi}_t \geq 3\Phi(V)/4 \geq 3\epsilon/4$. Thus

$$\#\text{ of hypotheses drawn } = \sum_{t=1}^{T} \left(\frac{48}{\widehat{\rho}_t^2 \epsilon} + \frac{32}{\widehat{\rho}_t^2 \widehat{\Phi}_t} + \frac{40}{\widehat{\Phi}_t^2} \right) \log \frac{4}{\delta_0} \leq \sum_{t=1}^{T} \left(\frac{96}{\epsilon \widehat{\rho}_t^2} + \frac{72}{\Phi(V)^2} \right) \log \frac{4}{\delta_0}$$

Given $\widehat{\rho}_t = 1/2^t$ and $T \leq 2 + \log 1/\rho$, we have

$$\sum_{t=1}^{T} \frac{1}{\widehat{\rho}_t^2} \;=\; \sum_{t=1}^{T} 2^{2t} \;=\; \frac{4}{3}(2^{2T} - 1) \;\leq\; \frac{4}{3} 2^{4+2\log 1/\rho} \;\leq\; \frac{22}{\rho^2}.$$

Plugging in $\delta_0 \leq \frac{\delta}{2m(2+\log(1/\rho))}$, we recover the theorem statement. □

```
Prior distribution $\pi$ over hypothesis class $\mathcal{H}$
Initial version space: $V = \mathcal{H}$
While $\frac{1}{n}\psi(E) \geq \frac{3\epsilon}{4}$ for $E \sim (\pi|_V)^{2\times n}$:
    • Draw $m$ data points $\mathbf{x} = (x_1, \ldots, x_m)$
    • Query point $x_i = \textsc{select}(V, \mathbf{x})$ and set $V$ to be consistent with the result
Return last version space $V$ in the form of the queried points $(x_1, h^*(x_1)), \ldots, (x_K, h^*(x_K))$
```

Figure 7.2. DBAL algorithm.

7.4.2 Active Learning Strategy

Using the SELECT procedure as a subroutine, Algorithm **??**, henceforth DBAL for Diameter-based Active Learning, is our active learning strategy. Given a hypothesis class with average splitting index $(\rho, \epsilon/2, \tau)$, DBAL queries data points provided by SELECT until it is confident $\Phi(V) < \epsilon$.

Denote by V_t the version space in the t-th round of DBAL. The following lemma, which is proven in the appendix, demonstrates that the halting condition (that is, $\psi(E) < 3\epsilon n/4$, where E consists of n pairs sampled from $(\pi|_V)^2$) guarantees that with high probability DBAL stops when $\Phi(V_t)$ is small.

Lemma 7.7. *The following holds for DBAL:*

(a) Suppose that for all $t = 1, 2, \ldots, K$ that $\Phi(V_t) > \epsilon$. Then the probability that the termination condition is ever true for any of those rounds is bounded above by $K \exp\left(-\frac{\epsilon n}{32}\right)$.

(b) Suppose that for some $t = 1, 2, \ldots, K$ that $\Phi(V_t) \leq \epsilon/2$. Then the probability that the termination condition is not true in that round is bounded above by $K \exp\left(-\frac{\epsilon n}{48}\right)$.

Given the guarantees on the SELECT procedure in Theorem 7.6 and on the termination condition provided by Lemma 7.7, we get the following theorem.

Theorem 7.8. *Suppose that $\mathcal{H}$ has average splitting index $(\rho, \epsilon/2, \tau)$. Then DBAL returns a version space V satisfying $\Phi(V) \leq \epsilon$ with probability at least $1 - \delta$ while using the following resources:*

(a) $K \leq \frac{8}{\rho}\left(\log\frac{2}{\epsilon} + 2\log\frac{1}{\pi(h^*)}\right)$ *rounds, with one label per round,*

(b) $m \leq \frac{1}{\tau}\log\frac{2K}{\delta}$ *unlabeled data points sampled per round, and*

(c) $n \leq O\Big(\left(\frac{1}{\epsilon\rho^2} + \frac{\log(1/\rho)}{\epsilon^2}\right)\left(\log\frac{mK}{\delta} + \log\log\frac{1}{\epsilon}\right)\Big)$ *hypotheses sampled per round.*

Proof. From definition of the average splitting index, if we draw $m = \frac{1}{\tau}\log\frac{2K}{\delta}$ unlabeled points per round, then with probability $1 - \delta/2$, each of the first K rounds will have at least one data point that ρ-average splits the current version space. In each such round, if the version space has average diameter at least $\epsilon/2$, then with probability $1 - \delta/4$ SELECT will return a data point that $\rho/8$-average splits the current version space while sampling no more than $n = O\left(\left(\frac{1}{\epsilon\rho^2} + \frac{1}{\epsilon^2}\log\frac{1}{\rho}\right)\log\frac{mK\log\frac{1}{\epsilon}}{\delta}\right)$ hypotheses per round by Theorem 7.6.

By Lemma 7.7, if the termination check uses $n' = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ hypotheses per round, then with probability $1 - \delta/4$ in the first K rounds the termination condition will never be true when the current version space has average diameter greater than ϵ and will certainly be true if the current version space has diameter less than $\epsilon/2$.

Thus it suffices to bound the number of rounds in which we can $\rho/8$-average split the version space before encountering a version space with average diameter $\epsilon/2$.

Since the version space is always consistent with the true hypothesis h^*, we will always have $\pi(V_t) \geq \pi(h^*)$. After $K = \frac{8}{\rho}\left(\log\frac{2}{\epsilon} + 2\log\frac{1}{\pi(h^*)}\right)$ rounds of $\rho/8$-average splitting, we have

$$\pi(h^*)^2\Phi(V_K) \;\leq\; \pi(V_K)^2\Phi(V_K) \;\leq\; \left(1 - \frac{\rho}{8}\right)^K \pi(V_0)^2\Phi(V_0) \;\leq\; \frac{\pi(h^*)^2\epsilon}{2}$$

Where we have used the fact that $\pi(V)^2\Phi(V) \leq 1$ for any set $V \subset \mathcal{H}$. Thus in the first K rounds, we must terminate with a version space with average diameter less than ϵ. □

7.5 Proof of Lemma 7.3

In this section, we give the proof of the following relationship between the original splitting index and our average splitting index.

Lemma 7.3. *Let π be a probability measure over a hypothesis class $\mathcal{H}$. If $\mathcal{H}$ has splitting index (ρ, ϵ, τ), then it has average splitting index $(\frac{\rho}{4\lceil\log(1/\epsilon)\rceil}, 2\epsilon, \tau)$.*

The first step in proving Lemma 7.3 is to relate the splitting index to our estimator $\psi(\cdot)$. Intuitively, splittability says that for any set of large edges there are many data points which remove a significant fraction of them. One may suspect this should imply that if a set of edges is large on average, then there should be many data points which remove a significant fraction of their weight. The following lemma confirms this suspicion.

Lemma 7.9. *Suppose that $V \subset \mathcal{H}$ has splitting index (ρ, ϵ, τ), and say E is a sequence of n hypothesis pairs from V satisfying $\psi(E) > 2n\epsilon$. If $x \sim \mathcal{D}$, then*

$$\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} \;\leq\; \left(1 - \frac{\rho}{4\lceil\log(1/\epsilon)\rceil}\right)\psi(E)$$

with probability at least τ.

Proof. Consider partitioning E as

$$\begin{aligned} E_0 &= \{\{h, h'\} \in E \,:\, d(h, h') < \epsilon\} \text{ and} \\ E_k &= \{\{h, h'\} \in E \,:\, d(h, h') \in [2^{k-1}\epsilon, 2^k\epsilon) \end{aligned}$$

for $k = 1, \ldots, K$ with $K = \lceil\log\frac{1}{\epsilon}\rceil$. Then $E_0, \ldots, E_K$ are all disjoint and their union is E. Define $E_{1:K} = \cup_{k=1}^K E_k$.

We first claim that $\psi(E_{1:K}) > \psi(E_0)$. This follows from the observation that because $\psi(E) \geq 2n\epsilon$ and each edge in E_0 has length less than ϵ, we must have

$$\psi(E_{1:K}) \;=\; \psi(E) - \psi(E_0) \;>\; 2n\epsilon - n\epsilon \;>\; \psi(E_0).$$

Next, observe that because each edge $\{h, h'\} \in E_k$ with $k \geq 1$ satisfies $d(h, h') \in [2^{k-1}\epsilon, 2^k\epsilon)$, we have

$$\psi(E_{1:K}) \;=\; \sum_{k=1}^{K} \sum_{\{h,h'\} \in E_k} d(h, h') \;\leq\; \sum_{k=1}^{K} 2^k \epsilon |E_k|.$$

Since there are only K summands on the right, at least one of these must be larger than $\psi(E_{1:K})/K$. Let k denote that index and let x be a point which ρ-splits E_k. Then we have

$$\begin{aligned}
\psi((E_{1:K})_x^+) \;&\leq\; \psi(E_{1:K}) - \psi(E_k \setminus (E_k)_x^+) \\
&\leq\; \psi(E_{1:K}) - \rho 2^{k-1}\epsilon |E_k| \\
&\leq\; \left(1 - \frac{\rho}{2K}\right) \psi(E_{1:K}).
\end{aligned}$$

Since $\psi(E_{1:K}) \geq \psi(E_0)$ and $\psi(E) = \psi(E_{1:K}) + \psi(E_0)$, we have

$$\psi(E_x^+) \;\leq\; \psi(E_0) + \left(1 - \frac{\rho}{2K}\right) \psi(E_{1:K}) \;\leq\; \left(1 - \frac{\rho}{4K}\right) \psi(E).$$

Symmetric arguments show the same holds for E_x^-. By the definition of splitting, the probability of drawing a point x which ρ-splits E_k is at least τ, giving us the lemma. □

With Lemma 7.9 in hand, we are now ready to prove Lemma 7.3.

Proof of Lemma 7.3. Let $V \subset \mathcal{H}$ such that $\Phi(V) > 2\epsilon$. Suppose that we draw n edges E i.i.d. from $\pi|_V$ and draw a data point $x \sim \mathcal{D}$. Then Hoeffding's inequality [55], combined with Lemma 7.4, tells us that there exist sequences $\epsilon_n, \delta_n \searrow 0$ such that with probability at least $1 - 3\delta_n$, the following hold simultaneously:

- $\Phi(V) - \epsilon_n \ \leq\ \frac{1}{n}\psi(E) \ \leq\ \Phi(V) + \epsilon_n$,
- $\frac{1}{n}\psi(E_x^+) \ \geq\ \frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+) - \epsilon_n$, and
- $\frac{1}{n}\psi(E_x^-) \ \geq\ \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-) - \epsilon_n$.

For ϵ_n small enough, we have that $\Phi(V) - \epsilon_n > 2\epsilon$. Combining the above with Lemma 7.9, we have with probability at least $\tau - 3\delta_n$,

$$\begin{aligned}
\max\left\{\frac{\pi(V_x^+)^2}{\pi(V)^2}\Phi(V_x^+), \frac{\pi(V_x^-)^2}{\pi(V)^2}\Phi(V_x^-)\right\} - \epsilon_n \ &\leq\ \frac{1}{n}\max\{\psi(E_x^+), \psi(E_x^-)\} \\
&\leq\ \left(1 - \frac{\rho}{4\lceil\log(1/\epsilon)\rceil}\right)\frac{\psi(E)}{n} \\
&\leq\ \left(1 - \frac{\rho}{4\lceil\log(1/\epsilon)\rceil}\right)(\Phi(V) + \epsilon_n).
\end{aligned}$$

By taking $n \to \infty$, we have $\epsilon_n, \delta_n \searrow 0$, giving us the lemma. □

7.6 Simulations

We compared DBAL against the baseline passive learner as well as two other generic active learning strategies: CAL and QBC. CAL proceeds by randomly sampling a data point and querying it if its label cannot be inferred from previously queried data points. QBC uses a prior distribution π and maintains a version space V. Given a randomly sampled data point x, QBC samples two hypotheses $h, h' \sim \pi|_V$ and queries x if $h(x) \neq h'(x)$.

We tested on two hypothesis classes: homogeneous, or through-the-origin, linear separators and k-sparse monotone disjunctions. In each of our simulations, we drew our target h^* from the prior distribution. After each query, we estimated the average diameter of the version space. We repeated each simulation several times and plotted the average performance of each algorithm.

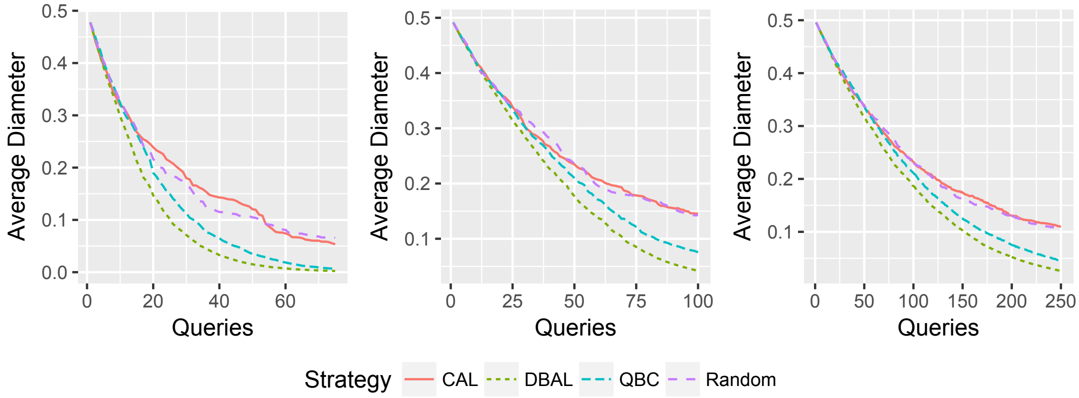


Figure 7.3. Simulation results on homogeneous linear separators. *Left*: $d = 10$. *Middle*: $d = 25$. *Right*: $d = 50$.

Homogeneous linear separators

The class of d-dimensional homogeneous linear separators can be identified with elements of the d-dimensional unit sphere. That is, a hypothesis $h \in \mathcal{S}^{d-1}$ acts on a data point $x \in \mathbb{R}^d$ via the sign of their inner product:

$$h(x) \ := \ \text{sign}(\langle h, x \rangle).$$

In our simulations, both the prior distribution and the data distribution are uniform over the unit sphere. Although there is no known method to exactly sample uniformly from the version space, Gilad-Bachrach et al. [44] demonstrated that samples generated by the hit-and-run Markov chain work well in practice. We adopted this approach for our sampling tasks.

Figure 7.3 shows the results of our simulations on homogeneous linear separators.

Sparse monotone disjunctions

A k-sparse monotone disjunction is a disjunction of k positive literals. Given a Boolean vector $x \in \{0, 1\}^n$, a monotone disjunction h classifies x as positive if and only if $x_i = 1$ for some positive literal i in h.

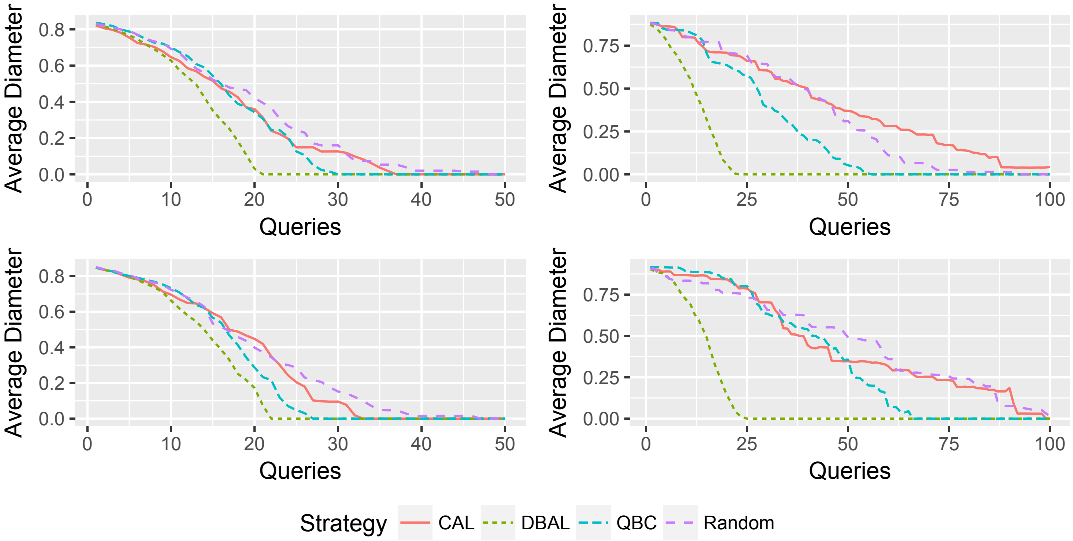


Figure 7.4. Simulation results on k-sparse monotone disjunctions. In all cases $k = 4$. *Top left*: $d = 75$, $p = 0.25$. *Top right*: $d = 75$, $p = 0.5$. *Bottom left*: $d = 100$, $p = 0.25$. *Bottom right*: $d = 100$, $p = 0.5$.

In our simulations, each data point is a vector whose coordinates are i.i.d. Bernoulli random variables with parameter p. The prior distribution is uniform over all k-sparse monotone disjunctions. When k is constant, it is possible to sample from the prior restricted to the version space in expected polynomial time using rejection sampling.

The results of our simulations on k-sparse monotone disjunctions are in Figure 7.4.

Chapter 7 contains material as it appears in "Diameter-based active learning." C. Tosh and S. Dasgupta. In International Conference of Machine Learning 2017. The dissertation author was the primary investigator.

Chapter 8

Structural query-by-committee

Chapter 7 presented an algorithm for the active learning setting in which the goal was to learn a low error classifier. In this chapter, we consider a broader interactive learning setting in which the goal is to find a generic structure, such as a clustering or hierarchy. We will see that this setting generalizes a variety of interactive learning scenarios.

8.1 Introduction

We introduce *interactive structure learning*, an abstract problem that encompasses many interactive learning tasks that have traditionally been studied in isolation, including active learning of binary classifiers, interactive clustering, interactive embedding, and active learning of structured output predictors. These problems include variants of both supervised and unsupervised tasks, and allow many different types of feedback, from binary labels to must-link/cannot-link constraints to similarity assessments to structured outputs. Despite these surface differences, they conform to a common template that allows them to be fruitfully unified.

In interactive structure learning, there is a space of items $\mathcal{X}$—for instance, an input space on which a classifier is to be learned, or points to cluster, or points to embed in a metric space—and the goal is to learn a *structure* on $\mathcal{X}$, chosen from a family $\mathcal{G}$. This set $\mathcal{G}$ could consist, for example, of all linear classifiers on $\mathcal{X}$, or all hierarchical clusterings of

$\mathcal{X}$, or all knowledge graphs on $\mathcal{X}$. There is a target structure $g^* \in \mathcal{G}$ and the hope is to get close to this target. This is achieved by combining a loss function or prior on $\mathcal{G}$ with interactive feedback from an expert.

We allow this interaction to be fairly general. In most interactive learning work, the dominant paradigm has been *question-answering*: the learner asks a question (like "what is the label of this point x?") and the expert provides the answer. We allow a more flexible protocol in which the learner provides a constant-sized *snapshot* of its current structure and asks whether it is correct ("does the clustering, restricted to these ten points, look right?"). If the snapshot is correct, the expert accepts it; otherwise, the expert fixes some part of it. This type of feedback, first studied in generality in [36], can be called *partial correction*. It is a strict generalization of question-answering, and as we explain in more detail below, it allows more intuitive interactions in many scenarios.

8.2 Interactive structure learning

The space of possible interactive learning schemes is large and mostly unexplored. We can get a sense of its diversity from a few examples. In *active learning* [80], for instance, the goal is to learn a classifier starting from a pool of unlabeled data. The machine adaptively decides which points it wants labeled, and an expert answers these queries as they arise. By focusing on informative points, the machine can often learn a good classifier using far fewer labels than would be needed in a passive setting.

Sometimes, the labels are complex structured objects, such as parse trees for sentences or segmentations of images. In such cases, providing an entire label is time-consuming, and it is easier if the machine simply suggests a label (such as a tree) and lets the expert either accept it or correct some particularly glaring fault in it. We can think of this as interaction with *partial correction*. It is more general than the *question-answering* usually assumed in active learning, and more convenient in many settings.

Interaction can also be used to augment *unsupervised* learning. Despite great improvements in algorithms for clustering, topic modeling, and so on, the outputs of these procedures are rarely perfectly aligned with the user's needs. Complex high-dimensional data can be organized in many different ways: should a collection of animals be clustered according to the Linnaean taxonomy, or their preferred habitats, or how cute they are? These alternatives are all legitimate, and it is impossible for an unsupervised method to magically guess what the user wants. But a modest amount of interaction can potentially overcome this problem of underspecification. For instance, the user can iteratively provide `must-link` and `cannot-link` constraints [86] to edit a flat clustering, or *triplet* constraints to edit a hierarchy [85].

These are just a few examples of the many types of interactive learning that have been investigated. The underlying tasks encompass problems of both supervised and unsupervised learning. The types of feedback range from triplets to partial labels to connectivity constraints. The querying strategies are also rich in variety. Our first goal is to provide a unifying framework in which this profusion of learning problems can be treated.

8.2.1 The space of structures

Let $\mathcal{X}$ be a set of data points. This could be a pool of unlabeled data to be used for active learning, or a set of points to be clustered, or an instance space on which a metric will be learned, or items on which a knowledge graph is to be constructed.

We wish to learn a *structure* on $\mathcal{X}$, chosen from a class $\mathcal{G}$. This could, for instance, be the set of all labelings of $\mathcal{X}$ consistent with a function class $\mathcal{F}$ of classifiers (binary, multiclass, or with complex structured labels), or the set of all partitions of $\mathcal{X}$, or the set of all metrics on $\mathcal{X}$. Of these structures, there is some target $g^* \in \mathcal{G}$ that we wish to attain.

Although interaction will help choose a structure, it is unreasonable to expect that interaction alone could be an adequate basis for this choice. For instance, pinpointing a particular clustering over n points requires $\Omega(n)$ must-link/cannot-link constraints, which is an excessive amount of interaction when n is large.

To bridge this gap, we need a prior or a loss function over structures. For instance, if $\mathcal{G}$ consists of flat k-clusterings, then we may prefer clusterings with low k-means cost. If $\mathcal{G}$ consists of linear separators, then we may prefer functions with small norm $\|g\|$. In the absence of interaction, the machine would simply pick the structure that optimizes the prior or cost function. In this paper, we assume that this preference is encoded as a prior distribution π over $\mathcal{G}$.

We emphasize that although we have adopted a Bayesian formulation, there is no assumption that the target structure g^* is actually drawn from the prior.

8.2.2 Feedback

We consider schemes in which each individual round of interaction is not expected to take too long. This means, for instance, that the expert cannot be shown an entire clustering, of unrestricted size, and asked to comment upon it. Instead, he or she can only be given a small *snapshot* of the clustering, such as its restriction to 10 elements. The feedback on this snapshot will be either be to accept it, or to provide some constraint that fixes part of it.

In order for this approach to work, it is essential that structures be *locally checkable*: that is, g corresponds to the target g^* if and only if every snapshot of g is satisfactory.

When g is a clustering, for instance, the snapshots could be restrictions of g to subsets $S \subseteq \mathcal{X}$ of some fixed size s. Technically, it is enough to take $s = 2$, which corresponds to asking the user questions of the form 'Do you agree with having `zebra` and `giraffe` in the same cluster?" From the viewpoint of human-computer interaction, it might be preferable to use larger subsets (like $s = 5$ or $s = 10$), with questions such as "Do you agree with the clustering $\{\texttt{zebra}, \texttt{giraffe}, \texttt{dolphin}\}, \{\texttt{whale}, \texttt{seal}\}$?" Larger substructures provide more context and are more likely to contain glaring faults that the user can easily fix (`dolphin` and `whale` must go together). In general, we can only expect the user to provide partial feedback in these cases, rather than fully correcting the substructure.

We now formalize the notion of a snapshot.

8.2.3 Snapshots

Perhaps the simplest type of snapshot of a structure g is *the restriction of g to a small number of points.* We start by discussing this case, and later present a more general definition.

Projections

For any $g \in \mathcal{G}$ and any subset $S \subseteq \mathcal{X}$ of size $s = O(1)$, let $g|_S$ be a suitable notion of the restriction of g to S, which we will sometimes call the *projection* of g onto S. For instance:

- $\mathcal{G}$ is a set of classifiers on $\mathcal{X}$.

  Then we can take $s = 1$. For any point $x \in \mathcal{X}$, we let $g|_x$ be $(x, g(x))$.

- $\mathcal{G}$ is a set of partitions (flat clusterings) of $\mathcal{X}$.

  For a set $S \subseteq \mathcal{X}$ of size $s \geq 2$, let $g|_S$ be the induced partition on just the points S.

- $\mathcal{G}$ is a set of hierarchical clusterings of $\mathcal{X}$.

  For any $s \geq 3$, and any set $S \subseteq \mathcal{X}$ of size s, let $g|_S$ be the restriction of the hierarchical clustering g to just the points S, that is, the induced hierarchy on s leaves.

- $\mathcal{G}$ is a set of metrics on $\mathcal{X}$.

For any $s \geq 2$ and any set $S \subseteq \mathcal{X}$ of size s, let $g|_S$ denote the $s \times s$ matrix of distances between points in S according to metric g.

As discussed earlier, from a human-computer interaction point of view, it will often be helpful to pick projections of size larger than the minimal possible s. For clusterings, for instance, any $s \geq 2$ satisfies local checkability, but human feedback might be more effective when $s = 10$ than when $s = 2$. Thus, in general, the queries made to the expert will consist of snapshots (projections of size $s = 10$, say) that can in turn be decomposed further into *atomic units* (projections of size 2).

Atomic decompositions of structures

Now we generalize the notion of projection to other types of snapshots and their atomic units.

We will take a functional view of the space of structures $\mathcal{G}$, in which each structure g is specified by its "answers" to a set of *atomic questions* $\mathcal{A}$. For instance, if $\mathcal{G}$ is the set of partitions of $\mathcal{X}$, then we can take $\mathcal{A} = \binom{\mathcal{X}}{2}$, with

$$g(\{x, x'\}) = \begin{cases} 1 & \text{if } g \text{ places } x, x' \text{ in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

The queries made during interaction can, in general, be composed of multiple atomic units, and feedback will be received on at least one of these atoms. Formally, let $\mathcal{Q}$ be the space of queries. In the partitioning example, this might be $\binom{\mathcal{X}}{10}$. The relationship between $\mathcal{Q}$ and $\mathcal{A}$ is captured by the following requirements:

- Each $q \in \mathcal{Q}$ can be decomposed as a set of atomic questions $A(q) \subseteq \mathcal{A}$. In the partitioning example, $A(q)$ is the set of all pairs in q.

- We will overload notation and write $g(q) = \{(a, g(a)) : a \in A(q)\}$.

- The user accepts $g(q)$ if and only if g satisfactorily answers every atomic question in q, that is, if and only if $g(a) = g^*(a)$ for all $a \in A(q)$.

To illustrate this notation, we briefly turn to the example of hierarchical clustering.

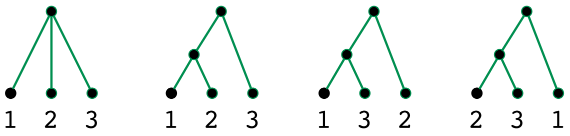
Example: hierarchical clustering

Suppose $\mathcal{G}$ is the space of hierarchical clusterings of $\mathcal{X}$ and the user has in mind a target hierarchy g^*.

A projection $g|_S$, the restriction of hierarchy g to leaves S, is correct if and only if it agrees exactly with $g^*|_S$. We can define the atomic questions to be projections of size 3, that is, $\mathcal{A} = \binom{\mathcal{X}}{3}$, and view any hierarchy g as a function:

$$g : \mathcal{A} \to \{\text{rooted trees with three leaves}\}.$$

Note that the hierarchy is fully specified by this function (to make this precise, we need to also fix some canonical ordering of the data points.) The right-hand set can be thought of as a set of possible labels, so that the learning problem resembles multiclass classification. There are four possible rooted trees with leaves $1, 2, 3$ and thus four labels:



The queries made by the machine can consist of larger projections, $\mathcal{Q} = \binom{\mathcal{X}}{s}$ for $s \geq 3$. Each such query q decomposes naturally into its constituent atomic questions: $A(q) = \{a \in \mathcal{A} : a \subseteq q\}$. For instance, if $s = 6$ then $|A(q)| = \binom{6}{3} = 20$.

8.2.4 Summary of framework

To summarize, interactive structure learning has two key components:

- A reduction to multiclass classifier learning.

We view each structure $g \in \mathcal{G}$ as a function on atomic questions $\mathcal{A}$. Thus, learning a good structure is equivalent to picking one whose labels $g(a)$ are correct.

- Feedback by partial correction.

  For practical reasons we consider broad queries, from a set $\mathcal{Q}$, where each query can be decomposed into atomic questions, allowing for partial corrections. This decomposition is given by the function $A : \mathcal{Q} \to 2^{\mathcal{A}}$.

The reduction to multiclass classification immediately suggests algorithms that can be used in the interactive setting. We are particular interested in *adaptive* querying, with the aim of finding a good structure with minimal interaction. Of the many schemes available for binary classifiers, one that appears to work well in practice and has good statistical properties is *query-by-committee* [81, 41]. It is thus a natural candidate to generalize to the broader problem of structure learning.

8.3 Structural QBC

Query-by-committee, as originally analyzed in [41], is an active learning algorithm for binary classification in the noiseless setting. It uses a prior probability distribution π over its classifiers and keeps track of the current version space, i.e. the classifiers consistent with the labeled data seen so far. At any given time, the next query is chosen as follows:

- Repeat:
  - Pick $x \in \mathcal{X}$ at random (e.g. from a pool of unlabeled data)
  - Pick classifiers h, h' at random from π restricted to the current version space
  - If $h(x) \neq h'(x)$: halt and take x as the point to query

In our setting, the feedback at time t is the answer y_t to some atomic question $a_t \in \mathcal{A}$, and we can define the resulting version space to be $\{g \in \mathcal{G} : g(a_{t'}) = y_{t'} \text{ for all } t' \leq t\}$. The

```
Prior $\pi$ over candidate structures $\mathcal{G}$
Distribution $\nu$ over query space $\mathcal{Q}$
Initial prior distribution over $\mathcal{G}$: $\pi_0 = \pi$

For $t = 1, 2, \ldots$:
    • Draw $g_t \sim \pi_{t-1}$
    • Repeat until the next query $q_t \in \mathcal{Q}$ has been chosen:
        – Draw $q \sim \nu$ and $g, g' \sim \pi_{t-1}$
        – With probability $d(g, g'; q)$: take $q_t = q$
    • Show user $q_t$ and $g_t(q_t)$
    • Expert corrects or confirms one or more atoms in $q_t$ by providing
      pairs $(a_t, y_t)$
    • Update posterior: $\pi_t(g) \propto \pi_{t-1}(g) \exp(-\beta \cdot \mathbf{1}(g(a_t) \neq y_t))$
      (do this for each feedback pair)
```

Figure 8.1. Structural QBC for $0 - 1$ loss.

immediate generalization of QBC would involve picking a query $q \in \mathcal{Q}$ at random, and then choosing it if g, g' sampled from π restricted to our version space happen to disagree on it. But this is unlikely to work well, because the answers to queries are no longer binary labels but mini-structures. As a result, g, g' are likely to disagree on minor details even when the version space is quite small, leading to excessive querying. To address this, we will use a more refined notion of the difference between $g(q)$ and $g'(q)$:

$$d(g, g'; q) \;=\; \frac{1}{|A(q)|} \sum_{a \in A(q)} \mathbf{1}[g(a) \neq g'(a)].$$

In words, this is the fraction of atomic subquestions of q on which g and g' disagree. It is a value between 0 and 1, where higher values mean that $g(q)$ differs significantly from $g'(q)$. Then we will query q with probability $d(g, g'; q)$.

8.3.1 Accommodating noisy feedback

We are interested in the noisy setting, where the user's feedback may occasionally be inconsistent with the target structure. In this case, the notion of a version space is less clear-cut. Our proposed modification is very simple: the feedback at time t, say (a_t, y_t), causes the posterior to be updated as follows:

$$\pi_t(g) \;\propto\; \pi_{t-1}(g)\exp(-\beta \cdot \mathbf{1}[g(a_t) \neq y_t]). \tag{8.1}$$

Here $\beta > 0$ is a constant that controls how aggressively errors are punished. In the noiseless setting, we can take $\beta = \infty$ and recover the original QBC update. Even with noise, however, we will demonstrate that this posterior update enjoys convergence guarantees. The full algorithm is shown in Figure 8.1.

8.3.2 Uncertainty and informative queries

What kinds of queries will structural QBC make? To answer this, we first quantify the *uncertainty* in the current posterior about a particular query or atom. For $a \in \mathcal{A}$ and $q \in \mathcal{Q}$ and any distribution π, write

$$\begin{aligned} u(a;\pi) &= \Pr_{g,g' \sim \pi}(g(a) \neq g'(a)) \\ u(q;\pi) &= \mathbb{E}_{g,g' \sim \pi}[d(g,g';q)] \;=\; \mathbb{E}_{a \sim \text{unif}(A(q))}[u(a;\pi)], \end{aligned}$$

where `unif` denotes the uniform distribution. These uncertainty values lie in the range $[0, 1]$.

The probability that a particular query $q \in \mathcal{Q}$ is chosen in round t by structural QBC is proportional to $\nu(q)u(q;\pi_{t-1})$. Thus, queries with higher uncertainty under the current posterior are more likely to be chosen. As the following lemma demonstrates, getting feedback on uncertain atoms leads to the elimination, or down-weighting in the case of noisy feedback, of many structures inconsistent with g^*.

Lemma 8.1. *Let π be a distribution over $\mathcal{G}$. For any $a \in \mathcal{A}$ and any answer y to a,*

$$\pi(\{g : g(a) \neq y\}) \;\geq\; \frac{1}{2} u(a; \pi).$$

Proof. Suppose the possible answers to a are $y_1, y_2, \ldots$, and that these have probabilities $p_1 \geq p_2 \geq \cdots$ respectively under π. That is $p_i = \pi(\{g : g(a) = y_i\})$. Then

$$u(a; \pi) \;=\; 1 - \sum_i p_i^2.$$

Note that $\pi(\{g : g(a) \neq y\})$ is smallest when $y = y_1$. Thus, we have

$$1 - \pi(\{g : g(a) \neq y\}) \;\leq\; p_1 \;\leq\; \sqrt{1 - u(a; \pi)} \;\leq\; 1 - \frac{1}{2} u(a; \pi),$$

Rearranging gives us the lemma. □

This gives some intuition for the query selection criterion of structural QBC, and will later be used in the proof of consistency.

8.3.3 General loss functions

The update rule for structural QBC, equation (8.1), results in a posterior of the form $\pi_t(g) \propto \pi(g) \exp(-\beta \cdot \#(\text{mistakes made by } g))$, which can in general be difficult to sample from. To address this, we consider a broader class of updates,

$$\pi_t(g) \;\propto\; \pi_{t-1}(g) \exp(-\beta \cdot \ell(g(a_t), y_t)), \tag{8.2}$$

where $\ell(\cdot, \cdot)$ is a general loss function. In the special case where $\mathcal{G}$ consists of linear functions and ℓ is convex, the resulting posterior is a log-concave distribution, which allows for efficient sampling [70]. We will show that this update also enjoys nice theoretical properties, albeit under different noise conditions.

```
Prior $\pi$ over candidate structures $\mathcal{G}$
Distribution $\nu$ over query space $\mathcal{Q}$
Initial prior distribution over $\mathcal{G}$: $\pi_0 = \pi$

For $t = 1, 2, \ldots$:
    • Draw $g_t \sim \pi_{t-1}$
    • Repeat until the next query $q_t \in \mathcal{Q}$ has been chosen:
        – Draw $q \sim \nu$ and $g, g' \sim \pi_{t-1}$
        – With probability $d^2(g, g'; q)$: take $q_t = q$
    • Show user $q_t$ and $g_t(q_t)$
    • Expert corrects or confirms one or more atoms in $q_t$ by providing
      pairs $(a_t, y_t)$
    • Update posterior: $\pi_t(g) \propto \pi_{t-1}(g) \exp(-\beta \cdot \ell(g(a_t), y_t))$
      (do this for each feedback pair)
```

Figure 8.2. Structural QBC for general loss functions.

To formally specify the setting, let $\mathcal{Y}$ be the space of answers to atomic questions $\mathcal{A}$, and suppose that structures in $\mathcal{G}$ generate values in some possibly different prediction space $\mathcal{Z} \subseteq \mathbb{R}^d$. That is, we view each $g \in \mathcal{G}$ as a function $g : \mathcal{A} \to \mathcal{Z}$, and any output $z \in \mathcal{Z}$ gets translated to some prediction in $\mathcal{Y}$. The loss associated with predicting z when the actual answer is y is denoted $\ell(z, y)$. Here are some examples:

- $0 - 1$ loss. $\mathcal{Z} = \mathcal{Y}$ and $\ell(z, y) = \mathbf{1}(y \neq z)$.

- Logistic loss. $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Z} = [-B, B]$ for $B > 0$, and $\ell(z, y) = \ln(1 + e^{-yz})$.

- Squared loss. $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Z} = [-B, B]$, and $\ell(z, y) = (y - z)^2$.

When moving from a discrete to a continuous prediction space, it becomes very possible that the predictions, on a particular atomic question, made by two randomly chosen structures will be close but not perfectly aligned. Thus, instead of checking strict equality of these predictions, we need to modify our querying strategy to take into account the distance between them. To this end, we will use the normalized average squared Euclidean distance:

$$d^2(g, g'; q) \;=\; \frac{1}{|A(q)|} \sum_{a \in A(q)} \frac{\|g(a) - g'(a)\|^2}{D}$$

where $D = \max_{a \in \mathcal{A}} \max_{g, g' \in \mathcal{G}} \|g(a) - g'(a)\|^2$. Note that $d^2(g, g'; q)$ is a value between 0 and 1, and thus we can treat it as a probability, similar to how we used $d(g, g'; q)$ in the 0-1 loss setting. The full algorithm is shown in Figure 8.2.

In the 0-1 loss setting, we saw that structural QBC chooses queries proportional to their uncertainty. What queries will structural QBC make in the general loss setting? Define the variance of $a \in \mathcal{A}$ under distribution π as

$$\mathrm{var}(a; \pi) = \sum_{g \in \mathcal{G}} \pi(g) \, \|g(a) - \mathbb{E}_{g' \sim \pi}[(g'(a))]\|^2 = \frac{1}{2} \sum_{g, g' \in \mathcal{G}} \pi(g) \, \pi(g') \, \|g(a) - g'(a)\|^2$$

and define the variance of a query $q \in \mathcal{Q}$ as the average variance of its constituent atoms,

$$\mathrm{var}(q; \pi) = \mathbb{E}_{a \sim \mathrm{unif}(A(q))}[\mathrm{var}(a; \pi)] = \frac{1}{|A(q)|} \sum_{a \in A(q)} \mathrm{var}(a; \pi).$$

Then it is not hard to see that the probability that structural QBC chooses $q \in \mathcal{Q}$ at step t is proportional to $\nu(q)\mathrm{var}(q; \pi_{t-1})$.

8.4 Kernelizing structural QBC

Consider the special case where $\mathcal{G}$ consists of linear functions, i.e. $\mathcal{G} = \{g_w(x) = \langle x, w \rangle \;:\; w \in \mathbb{R}^d\}$. As mentioned above, when the loss function is convex, the posteriors we encounter are log-concave, and thus efficiently samplable. But what if we want a more expressive class than linear functions? To address this, we resort to kernels.

Gilad-Bachrach et al. [44] investigated the use of kernels in QBC. In particular, they observed that to run QBC, we need not actually sample from the prior restricted to the current version space. Rather, given a candidate query x, it is enough to be able to sample from the distribution this posterior induces over the labelings of x. Although their work was in the realizable binary setting, this observation readily applies to our noisy structural setting.

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a *feature mapping.* Given a prior π over $\mathbb{R}^d$, the posterior after observing $(x_1, y_1), \cdots, (x_t, y_t)$ becomes

$$\pi_t(g_w) \propto \pi(g_w) \exp\left(-\beta \sum_{i=1}^{t} \ell(\langle \phi(x_i), w \rangle, y_i)\right).$$

A particularly interesting case is when $\ell(\cdot, \cdot)$ is the squared-loss and π is Gaussian. In this case, we will show that the predictions of the posterior are distributed according to a univariate Gaussian distribution with efficiently computable mean and variance. To show this, we first observe that the posterior is a multivariate Gaussian.

Lemma 8.2. *Suppose $\pi = N(0, \sigma_o^2 I_d)$, $\ell(\cdot, \cdot)$ is the squared-loss, and we have observed $(x_1, y_1), \cdots, (x_t, y_t)$. If $\Phi \in \mathbb{R}^{t \times d}$ denotes the matrix*

$$\Phi = \begin{bmatrix} — & \phi(x_1) & — \\ — & \phi(x_2) & — \\ & \vdots & \\ — & \phi(x_t) & — \end{bmatrix}.$$

then π_t is $N(\widehat{\mu}, \widehat{\Sigma})$ where $\widehat{\Sigma} = \left(2\beta\Phi^T\Phi + \frac{1}{\sigma_o^2} I_d\right)^{-1}$ and $\widehat{\mu} = 2\beta\widehat{\Sigma}\Phi^T y$.

We defer the proof of Lemma 8.2 to the appendix. Since π_t is a multivariate

Gaussian, we know that if $w \sim \pi_t$ and $v \in \mathbb{R}^d$ then $\langle w, v\rangle$ is distributed according to $N(\mu, \sigma^2)$ where

$$\begin{aligned}
\mu &= v^T\widehat{\mu} = 2\beta v^T\widehat{\Sigma}\Phi^T y \\
\sigma^2 &= v^T\widehat{\Sigma}v = v^T\left(2\beta\Phi^T\Phi + \frac{1}{\sigma_o^2}I_d\right)^{-1} v
\end{aligned}$$

Unfortunately, directly computing μ and σ^2 in the forms above requires expanding out the feature mappings, which is undesirable. However, the following theorem, known as the Woodbury Matrix Identity [53, Exercise 13.9], allows us to rewrite these terms in a form only involving inner products of the feature vectors.

Theorem 8.3 (Woodbury Matrix Identity)**.** *Let T, W, U, V be matrices of the appropriate sizes. Then*

$$(T + UW^{-1}V)^{-1} = T^{-1} - T^{-1}U(W + VT^{-1}U)^{-1}VT^{-1}.$$

Theorem 8.3 implies that we can rewrite $\widehat{\Sigma}$ as

$$\begin{aligned}
\widehat{\Sigma} &= \left(\frac{1}{\sigma_o^2}I_d + \Phi^T(2\beta I_n)\Phi\right)^{-1} \\
&= \sigma_o^2 I_d - \sigma_o^2 I_d\Phi^T\left(\frac{1}{2\beta}I_n + \Phi(\sigma_o^2 I_d)\Phi^T\right)^{-1}\Phi(\sigma_o^2 I_d) \\
&= \sigma_o^2\left(I_d - \Phi^T\left(\frac{1}{2\sigma_o^2\beta}I_n + \Phi\Phi^T\right)^{-1}\Phi\right) \\
&= \sigma_o^2\left(I_d - \Phi^T\Sigma_0\Phi\right)
\end{aligned}$$

where $\Sigma_o = \left(\frac{1}{2\sigma_o^2\beta}I_t + \Phi\Phi^T\right)^{-1}$. With this observation in hand, we have the following

Lemma 8.4. *Suppose the assumptions of Lemma 8.2 hold. If $g_w \sim \pi_t$, then $\langle w, \phi(x)\rangle$ is distributed according to $N(\mu, \sigma^2)$ where*

$$\mu \;=\; 2\sigma_o^2\beta\kappa^T\left(I_t - \Sigma_o K\right)y$$

$$\sigma^2 \;=\; \sigma_o^2 \left(\phi(x)^T\phi(x) - \kappa^T \Sigma_o \kappa\right)$$

where $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$, $\kappa_i = \langle \phi(x_i), \phi(x) \rangle$, *and* $\Sigma_o = \left(\frac{1}{2\sigma^2\beta} I_t + K\right)^{-1}$.

The important observation here is that all the quantities involving the feature mapping in Lemma 8.4 are inner products. Thus we never need to explicitly construct any feature vectors.

8.5 Consistency of structural QBC

In this section, we look at a typical setting in which there is a finite but possibly very large pool of candidate questions $\mathcal{Q}$, and thus the space of structures $\mathcal{G}$ is effectively finite. Let $g^* \in \mathcal{G}$ be the target structure, as before. Our goal in this setting is to demonstrate the *consistency* of structural QBC, meaning that

$$\lim_{t \to \infty} \pi_t(g^*) \;=\; 1$$

almost surely. To do so, we first formalize our setting. Note that the random outcomes during time step t of structural QBC consist of:

- the query q_t;
- the atomic question $a_t \in A(q_t)$ that the expert chooses to answer (pick one at random if the expert answers several of them); and
- the response y_t to a_t.

Let $\mathcal{F}_t$ denote the sigma-field of all outcomes up to, and including, time t. We begin with the special case of structural QBC under the 0-1 loss.

8.5.1 Consistency under 0-1 loss

In order to prove consistency, we will have to make some assumptions about the feedback we receive from a user. For any query $q \in \mathcal{Q}$ and any atomic question $a \in A(q)$, let $\eta(y|a, q)$ denote the conditional probability that the user answers y to atomic question a, in the context of query q. Our first assumption is that the most likely answer is $g^*(a)$.

Assumption 8.1. *There exists* $0 < \lambda \leq 1$ *such that* $\eta(g^*(a)|a, q) - \eta(y|a, q) \geq \lambda$ *for all* $q \in \mathcal{Q}$ *and* $a \in A(q)$ *and all* $y \neq g^*(a)$.

(We will use the convention $\lambda = 1$ for the noiseless setting.) In the learning literature, Assumption 8.1 is known as Massart's bounded noise condition [7]. As an example, suppose that there are 11 possible answers to an atom. A user that answers correctly with probability 0.10 and provides every other incorrect answer with probability 0.09 would satisfy Assumption 8.1 with $\lambda = 0.01$. Thus, Assumption 8.1 allows for users who are prone to mistakes but not inherently biased to a particular incorrect answer.

The following lemma demonstrates that under Assumption 8.1, the posterior probability of g^* increases in expectation with each query, as long as the β parameter of the update rule in equation (8.1) is small enough relative to λ.

Lemma 8.5. *Fix any* t, *and suppose the expert provides an answer to atomic question* $a_t \in A(q_t)$ *at time* t. *Let* $\gamma_t = \pi_{t-1}(\{g \in \mathcal{G} : g(a_t) = g^*(a_t)\})$. *Define* Δ_t *by:*

$$\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}, q_t, a_t\right] \;=\; (1 - \Delta_t)\frac{1}{\pi_{t-1}(g^*)},$$

Under Assumption 8.1, Δ_t *can be lower-bounded as follows:*

(a) If $\lambda = 1$ *(noiseless setting),* $\Delta_t \geq (1 - \gamma_t)(1 - e^{-\beta})$.

(b) For any $0 < \lambda \leq 1$, *if* $\beta \leq \lambda/2$, *then* $\Delta_t \geq \beta\lambda(1 - \gamma_t)/2$.

Proof. Let $y_1, y_2, \ldots$ denote the possible answers to a_t, and set $p_j = \eta(y_j | a_t, q_t)$ be the probability that the labeler answers y_j. Without loss of generality, suppose $p_1 \geq p_2 \geq \cdots$, so that (under Assumption 8.1) $g^*(a_t) = y_1$ and $p_1 - p_2 \geq \lambda$.

Further, define $\mathcal{G}_j = \{g \in \mathcal{G} : g(a_t) = y_j\}$. Thus $\gamma_t = \pi_{t-1}(\mathcal{G}_1)$. By averaging over the expert's possible responses, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}, q_t, a_t\right] \\
&= p_1 \frac{\pi_{t-1}(\mathcal{G}_1) + e^{-\beta}(1 - \pi_{t-1}(\mathcal{G}_1))}{\pi_{t-1}(g^*)} + \sum_{j>1} p_j \frac{\pi_{t-1}(\mathcal{G}_j) + e^{-\beta}(1 - \pi_{t-1}(\mathcal{G}_j))}{e^{-\beta}\pi_{t-1}(g^*)} \\
&= \frac{1}{\pi_{t-1}(g^*)}\left(p_1(\pi_{t-1}(\mathcal{G}_1) + e^{-\beta}(1 - \pi_{t-1}(\mathcal{G}_1)) + \sum_{j>1} p_j(e^{\beta}\pi_{t-1}(\mathcal{G}_j) + (1 - \pi_{t-1}(\mathcal{G}_j)))\right) \\
&= \frac{1}{\pi_{t-1}(g^*)}\left(p_1(1 - (1 - \pi_{t-1}(\mathcal{G}_1))(1 - e^{-\beta})) + \sum_{j>1} p_j(1 + (e^{\beta} - 1)\pi_{t-1}(\mathcal{G}_j))\right) \\
&= \frac{1}{\pi_{t-1}(g^*)} - \frac{1}{\pi_{t-1}(g^*)}\left(p_1(1 - \pi_{t-1}(\mathcal{G}_1))(1 - e^{-\beta}) - \sum_{j>1} p_j(e^{\beta} - 1)\pi_{t-1}(\mathcal{G}_j)\right).
\end{aligned}$$

Setting the parenthesized term to Δ_t, we have

$$\begin{aligned}
\Delta_t &\geq p_1(1 - \pi_{t-1}(\mathcal{G}_1))(1 - e^{-\beta}) - p_2(e^{\beta} - 1)\sum_{j>2} \pi_{t-1}(\mathcal{G}_j) \\
&= \left(p_1(1 - e^{-\beta}) - p_2(e^{\beta} - 1)\right)(1 - \gamma_t).
\end{aligned}$$

This yields (a) in the lemma statement. For (b), using the inequalities $e^{\beta} \leq 1 + \beta + \beta^2$ and $e^{-\beta} \leq 1 - \beta + \beta^2$ for $0 \leq \beta \leq 1$, we get

$$\begin{aligned}
\Delta_t &\geq \left(p_1(\beta - \beta^2) - p_2(\beta + \beta^2)\right)(1 - \gamma_t) \\
&\geq \beta\left((p_1 - p_2) - \beta(p_1 + p_2)\right)(1 - \gamma_t) \;\geq\; \beta(\lambda - \beta)(1 - \gamma_t).
\end{aligned}$$

Taking $\beta \leq \lambda/2$ completes the proof. □

To understand the requirement $\beta = O(\lambda)$, consider an atomic question on which there are just two possible labels, 1 and 2, and the expert chooses these with probabilities p_1 and p_2, respectively. If the correct answer according to g^* is 1, then $p_1 \geq p_2 + \lambda$ under Assumption 8.1. Let $\mathcal{G}_2$ denote structures that answer 2.

- With probability p_1, the expert answers 1, and the posterior mass of $\mathcal{G}_2$ is effectively multiplied by $e^{-\beta}$.

- With probability p_2, the expert answers 2, and the posterior mass of $\mathcal{G}_2$ is effectively multiplied by e^{β}.

The second outcome is clearly undesirable. In order for it to be counteracted, in expectation, by the first, β must be kept fairly small relative to p_1/p_2. The condition $\beta \leq \lambda/2$ is sufficient for this.

Thus, Lemma 8.5 asserts that structural QBC shrinks $1/\pi_t(g^*)$, in expectation, on every round: it corresponds to a random walk with a drift in the right direction. This drift is proportional to $\beta\lambda(1-\gamma_t)$, where γ_t is the probability mass, under the current posterior, of structures that agree with g^* on the atom a_t.

Lemma 8.5 does not, in itself, imply consistency. It is quite possible for $1/\pi_t(g^*)$ to keep shrinking but not converge to 1. Imagine, for instance, that the input space has two parts to it, and we keep improving on one of them but not the other. What we need is, first, to ensure that the queries q_t capture some portion of the uncertainty in the current posterior, and second, that the user chooses an atom that is at least slightly informative. The first condition is assured by the SQBC querying strategy. For the second, we need an assumption.

Assumption 8.2. *There is some minimum probability $p_o > 0$ for which the following holds. If the user is presented with a query q and a structure $g \in \mathcal{G}$ such that $g(q) \neq g^*(q)$, the user will provide feedback on some $a \in A(q)$ such that $g(a) \neq g^*(a)$ with probability at least p_o.*

To understand this, note that the interface allows the user to either correct a mistake in $g(q)$ or to corroborate part of it that is correct. The assumption asserts that the former occurs at least some fraction of the time, in situations where $g(q)$ is not perfect. It is one way of avoiding scenarios in which a user never provides feedback on a particular atom a. In such a pathological case, we might not be able to recover $g^*(a)$, and thus our posterior will always put some probability mass on structures that disagree with g^* on a.

The following lemma, whose proof is deferred to the appendix, gives lower bounds on the quantity $1 - \gamma_t$ under Assumption 8.2.

Lemma 8.6. *Suppose that $\mathcal{G}$ is finite and the user's feedback obeys Assumption 8.2. Then there exists a constant $c > 0$ such that for every round t*

$$\mathbb{E}[1 - \gamma_t \,|\, \mathcal{F}_{t-1}] \;\geq\; c\,\pi_{t-1}(g^*)^2(1 - \pi_{t-1}(g^*))^2$$

where $\gamma_t = \pi_{t-1}(\{g \in \mathcal{G} : g(a_t) = g^*(a_t)\})$, a_t is the atom the user provides feedback on, and the expectation is taken over the randomness in structural QBC and the user's response.

Together, Lemmas 8.5 and 8.6 show that the sequence $1/\pi_t(g^*)$ is a positive supermartingale that decreases in expectation at each round by an amount that depends on $\pi_t(g^*)$. The following lemma gives us a condition under which such stochastic processes can be guaranteed to converge to 1.

Lemma 8.7. *Suppose that there exists a continuous, non-negative function $f : [0, 1] \to \mathbb{R}_{\geq 0}$ such that $f(1) = 0$ and $f(x) > 0$ for all $x \in (0, 1)$. If for each $t \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}\right] \;\leq\; \frac{1}{\pi_{t-1}(g^*)} - f(\pi_{t-1}(g^*))$$

then $\pi_t(g^*) \to 1$ almost surely.

Proof. Let $X_t = \pi_t(g^*)$. By assumption, $\frac{1}{X_t}$ is a positive supermartingale, which implies $\lim_{t\to\infty} \frac{1}{X_t} = \frac{1}{X}$ exists and is finite with probability one. By the Continuous Mapping

Theorem, this implies $\lim_{t\to\infty} X_t = X$ and is non-zero with probability one. On the other hand, we have by the law of total expectation

$$1 \;\leq\; \mathbb{E}\left[\frac{1}{X_T}\right] \;\leq\; \frac{1}{\pi(g^*)} - \sum_{t=0}^{T-1} \mathbb{E}[f(X_t)].$$

for all $T \in \mathbb{N}$, which implies that $\lim_{t\to\infty} \mathbb{E}[f(X_t)] = 0$. By Fatou's lemma and the Continuous Mapping Theorem, we have

$$0 \;=\; \lim_{t\to\infty} \mathbb{E}[f(X_t)] \;=\; \mathbb{E}\left[\lim_{t\to\infty} f(X_t)\right] \;=\; \mathbb{E}\left[f(X)\right].$$

Thus, $f(X) = 0$ with probability one. Since f has only two potential zeros at 0 and 1, and since $X > 0$ with probability one, we conclude that $X = 1$ with probability one. □

As an immediate corollary, we have that structural QBC is consistent.

Theorem 8.8. *Suppose that $\mathcal{G}$ is finite, and Assumptions 8.1 and 8.2 hold. Then if structural QBC is run with a prior distribution π in which $\pi(g^*) > 0$, we have $\lim_{t\to\infty} \pi_t(g^*) = 1$ almost surely.*

Proof. Combining Lemmas 8.5 and 8.6, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}\right] &= (1 - \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}])\frac{1}{\pi_{t-1}(g^*)} \\
&\leq \left(1 - \frac{c\beta\lambda\pi_{t-1}(g^*)^2(1-\pi_{t-1}(g^*))^2}{2}\right)\frac{1}{\pi_{t-1}(g^*)} \\
&= \frac{1}{\pi_{t-1}(g^*)} - \frac{c\beta\lambda\pi_{t-1}(g^*)(1-\pi_{t-1}(g^*))^2}{2}
\end{aligned}$$

Now $f(x) = \frac{c\beta\lambda x(1-x)^2}{2}$ meets all of the conditions of Lemma 8.7, which concludes the proof. □

In Section 8.6 we provide rates of convergence.

8.5.2 Consistency under general losses

We now turn to analyzing structural QBC with general losses. As before, we will need to make some assumptions. The first is that the loss function is well-behaved.

Assumption 8.3. *The loss function is bounded,* $0 \leq \ell(z, y) \leq B$*, and Lipschitz in its first argument,*

$$\ell(z, y) - \ell(z', y) \;\leq\; C\|z - z'\|,$$

for some constants $B, C > 0$.

It is easily checked that this assumption holds for the three loss functions we mentioned earlier.

In the case of 0-1 loss, we assumed that for any atomic question a, the correct answer $g^*(a)$ would be given with higher probability than any incorrect answer. We now formulate an analogous assumption for the case of more general loss functions. Recall that $\eta(\cdot|a)$ is the conditional probability distribution over the user's answers to $a \in \mathcal{A}$ (we can also allow η to also depend upon the context q, as we did before; here we drop the dependence for notational convenience). The expected loss incurred by $z \in \mathcal{Z}$ on this question is thus

$$L(z, a) \;=\; \sum_y \eta(y|a)\,\ell(z, y).$$

We will require that for any atomic question a, this expected loss is minimized when $z = g^*(a)$, and predicting any other z results in excess expected loss that grows with the distance between z and $g^*(a)$.

Assumption 8.4. *There exists a constant* $\lambda > 0$ *such that for any atomic question* $a \in \mathcal{A}$ *and any* $z \in \mathcal{Z}$*,*

$$L(z, a) - L(g^*(a), a) \;\geq\; \lambda\|z - g^*(a)\|^2.$$

Let's look at what this assumption implies in some concrete settings.

- $0-1$ loss with $\mathcal{Y} = \mathcal{Z} = \{0,1\}$. For any $z \in \{0,1\}$, we have

  $$L(z,a) = \sum_y \eta(y|a)\ell(z,y) = 1 - \eta(z|a)$$

  and thus Assumption 8.4 is equivalent to requiring

  $$\eta(g^*(a)|a) - \eta(z|a) \geq \lambda$$

  for all $a \in \mathcal{A}$ and $z \neq g^*(a)$. This is identical to the earlier Assumption 8.1.

- Squared loss with $\mathcal{Y} = \{-1,1\}$ and $\mathcal{Z} \subset \mathbb{R}$. Assumption 8.4 requires that for any $a \in \mathcal{A}$,

  $$g^*(a) = \operatorname*{argmin}_z L(z,a) = \operatorname*{argmin}_z \sum_y \eta(y|a)(z-y)^2 = \operatorname*{argmin}_z \mathbb{E}[(z-y)^2|a] = \mathbb{E}[y|a],$$

  where the expectation is over the choice of y given a. If this holds, then for any z, by a standard bias-variance decomposition, $L(z,a) - L(g^*(a),a) = (z - g^*(a))^2$, so that $\lambda = 1$.

- Logistic loss with $\mathcal{Y} = \{-1,1\}$ and $\mathcal{Z} = [-B,B]$. Fix any $a \in A$, and write $p = \eta(1|a)$. Then

  $$L(z,a) = \sum_y \eta(y|a)\ell(z,y) = p\ln(1+e^{-z}) + (1-p)\ln(1+e^z).$$

  This is minimized by $g^*(a) = \ln p - \ln(1-p)$, and $L(g^*(a),a)$ is then the entropy of a coin with bias p. Now pick any other value of z, and define $q = 1/(1+e^{-z})$ so that $z = \ln q - \ln(1-q)$. A further calculation shows that $L(z,a) - L(g^*(a),a)$ is exactly the KL divergence $K(p,q)$. Using Pinsker's inequality to lower-bound this in terms of $(p-q)^2$, we then find that Assumption 8.4 is satisfied with

$$\lambda = 2\left(\frac{e^B}{(1+e^B)^2}\right)^2.$$

From these examples, it is clear that requiring $g^*(a)$ to be the minimizer of $L(z,a)$ is plausible if $\mathcal{Z}$ is a discrete space but much less so if $\mathcal{Z}$ is continuous. In general, we can only hope that this holds approximately. With this caveat in mind, we stick with Assumption 8.4 as a useful but idealized mathematical abstraction.

Given these notions, and recalling our earlier definitions of $\mathrm{var}(a;\pi)$ and $\mathrm{var}(q;\pi)$, we have the following general loss analogue of Lemma 8.5.

Lemma 8.9. *Suppose Assumptions 8.3 and 8.4 hold. Take $\beta \leq \min(\lambda/(2C^2), 1/B)$. Suppose that at time t, the user provides feedback on an atomic question $a_t \in A(q_t)$ for which* $\mathrm{var}(a_t;\pi_{t-1}) \geq \gamma$*. Then*

$$\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}, q_t, a_t\right] \;=\; (1-\Delta_t)\frac{1}{\pi_{t-1}(g^*)},$$

where $\Delta_t \geq \beta\gamma\lambda/2$.

Proof. Plugging in the update rule for the posterior distribution, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}, q_t, a_t\right] &= \sum_y \eta(y|a_t)\frac{\sum_{g\in\mathcal{G}}\pi_{t-1}(g)\exp(-\beta\cdot\ell(g(a_t),y))}{\pi_{t-1}(g^*)\exp(-\beta\cdot\ell(g^*(a_t),y))}\\ &= \frac{1}{\pi_{t-1}(g^*)}\sum_y \eta(y|a_t)\sum_g \pi_{t-1}(g)\exp(-\beta\left[\ell(g(a_t),y)-\ell(g^*(a_t),y)\right]).\end{aligned}$$

To turn this expression into the form $(1-\Delta_t)\frac{1}{\pi_{t-1}(g^*)}$, define

$$\Delta_t = \sum_y \eta(y|a_t)\sum_g \pi_{t-1}(g)(1-\exp(-\beta\left[\ell(g(a_t),y)-\ell(g^*(a_t),y)\right])).$$

Since $\beta \leq 1/B$, and losses lie in $[0,B]$, we have that $\beta[\ell(g(a_t),y)-\ell(g^*(a_t),y)]$ lies in the range $[-1,1]$. Using the inequality $e^x \leq 1 + x + x^2$ for $-1 \leq x \leq 1$, we can lower bound Δ_t by

$$\begin{aligned}
&\sum_g \pi_{t-1}(g) \sum_y \eta(y|a_t)(\beta[\ell(g(a_t),y) - \ell(g^*(a_t),y)] - \beta^2[\ell(g(a_t),y) - \ell(g^*(a_t),y)]^2) \\
&= \sum_g \pi_{t-1}(g) \left(\beta[L(g(a_t),a_t) - L(g^*(a_t),a_t)] - \beta^2 \sum_y \eta(y|a_t)[\ell(g(a_t),y) - \ell(g^*(a_t),y)]^2 \right) \\
&\geq \sum_g \pi_{t-1}(g) \left(\beta\lambda \|g(a_t) - g^*(a_t)\|^2 - \beta^2 \sum_y \eta(y|a_t) C^2 \|g(a_t) - g^*(a_t)\|^2 \right) \\
&= (\beta\lambda - \beta^2 C^2) \sum_g \pi_{t-1}(g) \, \|g(a_t) - g^*(a_t)\|^2 \\
&\geq (\beta\lambda - \beta^2 C^2) \mathrm{var}(a_t; \pi_{t-1}) \;\geq\; \beta(\lambda - \beta C^2)\gamma,
\end{aligned}$$

where the second and third lines have used Assumptions 8.3 and 8.4. □

Similarly, we can also give a general loss analogue of Lemma 8.6.

Lemma 8.10. *Suppose $\mathcal{G}$ is finite and Assumption 8.2 holds. Then there exists a constant $c > 0$ such that for any round t*

$$\mathbb{E}[\mathrm{var}(a_t; \pi_{t-1}) \,|\, \mathcal{F}_{t-1}] \geq c\, \pi_{t-1}(g^*)^3 (1 - \pi_{t-1}(g^*))^2$$

where a_t is the atom the user provides feedback on and the expectation is taken over both the randomness of user's response and the randomness of structural QBC.

The proof of Lemma 8.10 is deferred to the appendix. With Lemmas 8.9 and 8.10 in hand, we get the consistency of general-loss structural QBC as a corollary.

Theorem 8.11. *Suppose $\mathcal{G}$ is finite and the user's feedback satisfies Assumptions 8.2, 8.3, and 8.4. If the general loss version of structural QBC is run with a prior distribution π in which $\pi(g^*) > 0$, then $\lim_{t\to\infty} \pi_t(g^*) = 1$ almost surely.*

8.6 Convergence rates

In this section, we bound the rate at which structural QBC's posterior concentrates on the target structure g^* under the 0-1 loss. Before we do so, we first introduce some concepts related to the informativeness of feedback.

8.6.1 Quantifying the informativeness of feedback

For an atomic question a and corresponding answer y, let $\mathcal{G}_{a,y} = \{g \in \mathcal{G} : g(a) = y\}$. Define the *shrinkage* of posterior π' due to atomic question a to be

$$S(\pi, a) = 1 - \max_{y} \pi(\mathcal{G}_{a,y})$$

and define the shrinkage of π' due to a query q to be the average shrinkage due to q's atoms

$$S(\pi, q) = \frac{1}{|A(q)|} \sum_{a \in A(q)} S(\pi, a).$$

This is very similar to the notion of *information gain* from the original query-by-committee analysis [41], as we explain in further detail at the end of this section. The reason for the modification is that it facilitates the generalizations introduced here: the multiclass setting and noisy feedback.

In Section 8.5, we saw that by making relatively weak assumptions on the specific atoms a user will provide feedback on when presented with a query, we can guarantee the consistency of structural QBC. To get meaningful convergence rates, however, we will need to make a stronger assumption. Specifically, we will require that when presented with a query q, the user provides feedback on an atom $a \in A(q)$ whose shrinkage is close to the average shrinkage of the atoms in $A(q)$.

Assumption 8.5. *When shown q_t and $g_t(q_t)$, the user provides feedback that satisfies*

$$\mathbb{E}[S(\pi_{t-1}, a_t)] \;\geq\; S(\pi_{t-1}, q_t)$$

where the expectation is taken over the randomness of the user's choice of a_t.

Later in this section, we characterize the shrinkage that might be expected in a few cases of interest. For the time being, we will think of it as one of the key quantities controlling the efficacy of active learning, and give rates of convergence in terms of it.

8.6.2 Rate of convergence

In this section, we will again assume $\mathcal{G}$ is finite. To keep things simple, we think of convergence as occurring when $\pi_t(g^*)$ exceeds some particular threshold τ (such as 1/2).

Theorem 8.12. *Suppose that $\mathcal{G}$ is finite, and that the expert's responses satisfy Assumptions 8.1 and 8.5. Suppose moreover that there are constants $0 < \tau, s_o < 1$ such that at any time t, if $\pi_t(g^*) \leq \tau$, the shrinkage of the next query is bounded below in expectation as*

$$\mathbb{E}[S(\pi_{t-1}, q_t)|\mathcal{F}_{t-1}] \;\geq\; s_o.$$

Pick any $0 < \delta < 1$. With probability at least $1 - \delta$, the number of rounds T of querying before $\pi_T(g^*) > \tau$ can be upper-bounded as follows.

$$T \;\leq\; \begin{cases} \frac{2}{s_o(1-e^{-\beta})} \max\left(\ln \frac{1}{\pi(g^*)}, \frac{4}{s_o(1-e^{-\beta})} \ln \frac{1}{\delta}\right) & \textit{if } \lambda = 1 \textit{ (noiseless case)} \\ \frac{4}{\beta\lambda s_o} \max\left(\ln \frac{1}{\pi(g^*)}, \frac{8e^{\beta}}{\beta\lambda s_o} \ln \frac{1}{\delta}\right) & \textit{for any } \lambda\textit{, if } \beta \leq \lambda/2 \end{cases}$$

Proof. We will spell out the argument for the noisy case; the other case is similar but slightly simpler. Define

$$R_t \;=\; 1 - \frac{\pi_{t-1}(g^*)}{\pi_t(g^*)}.$$

Using the random variable Δ_t from Lemma 8.5, we have

$$\begin{aligned}
\mathbb{E}[R_t|\mathcal{F}_{t-1}] &= 1 - \pi_{t-1}(g^*)\,\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}\right] \\
&= 1 - \pi_{t-1}(g^*)\,\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\pi_t(g^*)}\middle|\mathcal{F}_{t-1}, q_t, a_t\right]\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}[\Delta_t|\mathcal{F}_{t-1}] \\
&\geq \frac{1}{2}\beta\lambda\,\mathbb{E}[1-\gamma_t|\mathcal{F}_{t-1}] \\
&\geq \frac{1}{2}\beta\lambda\,\mathbb{E}[S(\pi_{t-1}, q_t)|\mathcal{F}_{t-1}] \;\geq\; \frac{1}{2}\beta\lambda s_o.
\end{aligned}$$

as long as $\pi_{t-1}(g^*) \leq \tau$ throughout.

Pick any time T before $\pi_T(g^*)$ exceeds τ. Then $\mathbb{E}[R_1 + \cdots + R_T] \geq \beta\lambda s_o T/2$. To show that this sum is concentrated around its expected value, we can use a martingale large deviation bound. First we check that each R_t is bounded. Since

$$e^{-\beta}\pi_{t-1}(g^*) \;\leq\; \pi_t(g^*) \;\leq\; e^{\beta}\pi_{t-1}(g^*),$$

it follows that R_t lies in an interval of size at most e^β. By the Azuma-Hoeffding inequality [55, 11], if T attains the value in the theorem statement, then

$$R_1 + \cdots + R_T \;>\; \frac{1}{2}\mathbb{E}[R_1 + \cdots + R_T] \;\geq\; \frac{1}{4}\beta\lambda s_o T \;\geq\; \ln\frac{1}{\pi(g^*)}.$$

with probability at least $1-\delta$. But this is not possible, since $R_1 + \cdots + R_T$ can be at most $\ln(1/\pi(g^*))$ by the chain of inequalities

$$1 \leq \frac{1}{\pi_T(g^*)} = (1-R_1)(1-R_2)\cdots(1-R_T)\frac{1}{\pi(g^*)} \leq \exp(-(R_1 + \cdots + R_T))\frac{1}{\pi(g^*)}. \quad \square$$

It is likely that the quadratic dependence on s_o, λ, and β can be reduced by a more careful large deviation bound.

8.6.3 Shrinkage and uncertainty

The active learning literature has introduced a variety of different complexity measures that attempt to capture the number of queries needed for learning. These include the *disagreement coefficient* [51] and the aforementioned *information gain* [41]. Although it is possible to give general bounds in terms of these quantities, it has proved quite difficult to compute these complexity measures for all but a few simple cases.

Our notion of shrinkage is a reformulation of the information gain that avoids assuming that the target structure is drawn from the prior distribution, and that accommodates scenarios beyond binary classification. By way of illustration, we will give an example of a simple situation in which the shrinkage can be characterized.

The following lemma demonstrates that under the structural QBC strategy, we can relate a query's shrinkage to its uncertainty.

Lemma 8.13. *Suppose the current posterior distribution is π_{t-1}. Under Assumption 8.5, the shrinkage S_t of the user's next response has expected value*

$$\mathbb{E}[S_t] \;\geq\; \frac{\mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})^2]}{\mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})]}.$$

Proof. Fix any query $q \in \mathcal{Q}$ and atom $a \in A(q)$. From Lemma 8.1, we have

$$S_t = \min_y \pi(\{g : g(a) \neq y\}) \geq u(a;\pi_t)/2.$$

Now, if the next query is q_t, then under Assumption 8.5,

$$S_t \;\geq\; \mathbb{E}_{a\sim\mathrm{unif}(A(q_t))}\frac{1}{2}u(a;\pi_{t-1}) \;=\; \frac{1}{2}u(q_t;\pi_{t-1}).$$

The lemma follows by taking expectation over $q_t \sim \nu$. □

We now turn to an example where we can bound the shrinkage.

Example: Partitioning a hypercube by axis-parallel cuts

Let $\mathcal{X}$ be the hypercube $[0,1]^p$, and consider axis-aligned bipartitions of $\mathcal{X}$. Any such partition is specified by a coordinate i and a value v, and yields clusters

$$\{x : x_i \leq v\} \text{ and } \{x : x_i > v\}.$$

Let π be the uniform distribution over such partitions $\mathcal{G} = [p] \times [0,1]$, so that $\pi(i,v) = 1/p$; and suppose that the data distribution is uniform over $\mathcal{X}$.

We will take atomic queries to be of the form $\{x, y\}$ for $x, y \in [0,1]^p$, where the answer is 1 if they lie in the same cluster, and 0 otherwise. The following lemma shows that the shrinkage of structural QBC queries is always constant in this setting.

Lemma 8.14. *Under Assumption 8.5, the shrinkage S_t of a user's feedback satisfies $\mathbb{E}[S_t] \geq 1/3$.*

Proof. Note that if we query $\{x, y\}$ and they are in separate clusters, the version space shrinks to the regions between x_i and y_i on each coordinate i; while if they are in different clusters, we get the complement. Either way, the resulting version space is isomorphic to the original $(\mathcal{G}, \pi)$, and hence this is the only case we need consider in computing uncertainty and shrinkage values.

For any $x, y \in [0,1]^p$, the probability that they are separated by a random draw from π is exactly

$$\sum_{i=1}^{p} \Pr(\text{cut coordinate is } i)\ |x_i - y_i| \ = \ \frac{\|x - y\|_1}{p}.$$

Thus the uncertainty on a query $\{x, y\}$ is

$$u(\{x, y\}) \;=\; \Pr_{g,g\sim\pi}(\text{exactly one of } g, g' \text{ separates } x,\, y) \;=\; 2\cdot\frac{\|x-y\|_1}{p}\left(1-\frac{\|x-y\|_1}{p}\right).$$

We will compute the expectation of this over $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_p)$ drawn uniformly at random from $[0,1]^p$.

First, a simple one-dimensional calculation shows that

$$\mathbb{E}[\|X-Y\|_1] \;=\; \sum_{i=1}^{p} \mathbb{E}|X_i - Y_i| \;=\; \frac{p}{3}.$$

Likewise,

$$\mathbb{E}[\|X-Y\|_1^2] \;=\; \sum_{i=1}^{p} \mathbb{E}|X_i - Y_i|^2 + \sum_{i\neq j}(\mathbb{E}|X_i - Y_i|)(\mathbb{E}|X_j - Y_j|) \;=\; \frac{p}{6} + \frac{p(p-1)}{9}.$$

Inserting these into the expression for uncertainty, we get

$$\mathbb{E}[u(\{X, Y\})] \;=\; 2\left(\frac{\mathbb{E}\|X-Y\|_1}{p} - \frac{\mathbb{E}\|X-Y\|_1^2}{p^2}\right) \;=\; \frac{4}{9} - \frac{1}{9p} \;\geq\; \frac{1}{3}.$$

We finish by invoking Lemma 8.13 and observing that

$$\mathbb{E}[S_t] \;\geq\; \mathbb{E}[u(\{X, Y\})^2]/\mathbb{E}[u(\{X, Y\})] \;\geq\; \mathbb{E}[u(\{X, Y\})]. \qquad \square$$

8.6.4 Relation to information gain

The original analysis of query-by-committee was specifically for active learning of binary classifiers and was based on the notion of *information gain* [41]. Suppose the current posterior distribution over classifiers is π, and that under this posterior, a specific query x has probability p of having a positive label and probability $1-p$ of having a negative label. The information gain here is the entropy of a coin with bias p,

$$I(\pi, x) \;=\; H(p).$$

In this same situation, the shrinkage is

$$S(\pi, x) \;=\; 1 - \max(p, 1-p).$$

These two quantities are related by a monotonic transformation. The analysis of QBC's query complexity assumes that the *expected* information gain, taken over the random choice of next query, is always bounded below by a constant. In the analysis presented in this section, we assumed the same of the expected shrinkage. In the case of binary classification, these two conditions coincide.

For instance, [41] showed that if the classifiers are homogeneous (through-the-origin) linear separators, and the data distribution is uniform over the unit sphere, then the expected information gain due to a label query is bounded below by a constant. This means that the same holds for the expected shrinkage.

8.7 Experiments

We now turn to our experiments with structural QBC in several applications. Before we do so, we first consider a way to speed up the query selection procedure.

8.7.1 Reducing the randomness in structural QBC

It is easy to see that the query selection procedure of structural QBC is a rejection sampler where each query q is chosen with probability proportional to $\nu(q)u(q;\pi_t)$ (in the case of the zero-one loss) or $\nu(q)\mathrm{var}(q;\pi_t)$ (for general losses). However, without knowing the normalization constant, it is possible for the rejection rate to be quite high, even when

```
Fixed set of queries $q_1, \ldots, q_m \in \mathcal{Q}$
Current distribution over $\mathcal{G}$: $\pi$
Initial shrinkage estimate: $\widehat{u}_o = 1/2$

For $t = 0, 1, 2, \ldots$:
    • Draw $g_1, g_1', \ldots, g_{n_t}, g_{n_t}' \sim \pi$
    • If there exists $q_j$ such that $\frac{1}{n_t} \sum_{i=1}^{n_t} d(g_i, g_i'; q_j) \geq \widehat{u}_t$ then **halt** and **return** $q_j$
    • Otherwise, let $\widehat{u}_{t+1} = \widehat{u}_t / 2$.
```

Figure 8.3. Robust query selection for structural QBC, under 0-1 loss.

there are many queries that have much higher uncertainty or variance than the rest. To circumvent this issue, we introduce a 'robust' version of structural QBC, wherein many candidate queries are sampled, and the query that has the highest uncertainty or variance is chosen.

In the zero-one loss case, we can estimate the uncertainty of a candidate query q by first drawing many pairs $g_1, g_1', \ldots, g_n, g_n' \sim \pi_t$ and then using the unbiased estimator

$$\widehat{u}(q; \pi_t) := \frac{1}{n} \sum_{i=1}^{n} d(g_i, g_i'; q).$$

By Hoeffding's inequality, this quantity concentrates tightly around the true uncertainty of q. Unfortunately, the number of structures we need to sample in order to identify the best candidate depends on the uncertainty of that candidate, which we do not know a priori. To circumvent this difficulty, we can start with an optimistic estimate of the largest uncertainty and then iteratively halve our estimate until we are confident that we have found a query with at least that much uncertainty. If the appropriate number of structures are sampled at each round, then it can be shown that this procedure terminates with a candidate query whose uncertainty is within a constant factor of the highest uncertainty (this is very similar to Lemma 7.5). The full algorithm for the zero-one loss case is shown in Figure 8.3.

In experiments, we have found that simply drawing a large number of candidate structures and choosing the query with the highest empirical uncertainty over these works quite well.

8.7.2 Clustering

In the case of flat clustering, queries can be taken to be restrictions of the current clustering restricted to a subset of the data points, and feedback can consist of must-link/cannot-link constraints over those data points. In each of our simulations, we ran the robust version of structured QBC where the candidate queries are subsets of size ten and feedback consists of a random correction to the proposed subset clustering. Below we describe the clustering models we used.

Mixture of Gaussians

Consider the following Bayesian generative model for a mixture of k spherical Gaussians.

- Weight vector $(w_1, \ldots, w_k)$ is drawn from a symmetric Dirichlet distribution with parameter $\alpha > 0$
- Means $\mu^{(1)}, \ldots, \mu^{(k)} \in \mathbb{R}^d$ are drawn i.i.d. from $N(\mu_o, \sigma_o^2 I_d)$
- For each data point $i = 1, \ldots, n$:
  - $z_i \in \{1, \ldots, k\}$ is drawn from Categorical$(w_1, \ldots, w_k)$
  - $x^{(i)} \in \mathbb{R}^d$ is drawn from $N(\mu_{z_i}, \sigma^2 I_d)$

Here, $x^{(1)}, \ldots, x^{(n)}$ are the observed data points; α, μ_o, σ_o^2, and σ^2 are known hyper-parameters; and the w's, μ's, and z's are the unobserved latent variables. Note that the z's induce a clustering over the data points. Our goal is to find the target clustering, with the Bayesian posterior distribution $\Pr(z \mid x)$ acting as our prior π over clusterings for structural QBC.

To sample from this posterior, we used the collapsed Gibbs sampler. Although this Markov chain may mix slowly, as shown in Chapter 5, we found that running structural QBC with these samples often led to fast convergence to the underlying clustering.

We ran experiments on the `wine` and `iris` datasets from the UCI machine learning repository [67]. In each of our experiments, we compared the robust structural QBC strategy (denoted 'Robust SQBC') against three baseline strategies:

- 'Random': feedback was provided on randomly sampled pairs of points.
- 'Random (correction)': feedback was a random correction to a proposed clustering of ten random points.
- 'Vanilla': no feedback at all, just an unconstrained Gibbs sampler.

The target clustering in each of the datasets was the one induced by the labels of the data points. Thus, in both of these datasets, the space of structures was clusterings containing at most 3 clusters.

For all strategies, we measured the clustering distance from the current clustering to the target clustering, i.e. the fraction of pairs of points the clusterings disagree on, and the log-posterior probability after every pass of the Gibbs sampler. For the feedback strategies, a constraint was added every 50 passes of the Gibbs sampler. The results are displayed in Figures 8.4.

As can be seen in both of these experiments, clustering error does not perfectly match log-posterior probability. In the `iris` experiments, for example, we see that the SQBC model does converge upon the target clustering but its log-posterior probability is still much lower than that of the clusterings found by the pure Gibbs sampler. Thus, even though the prior distribution does not put much weight upon the target, the SQBC strategy is still able to find the target clustering after a few rounds of interaction.

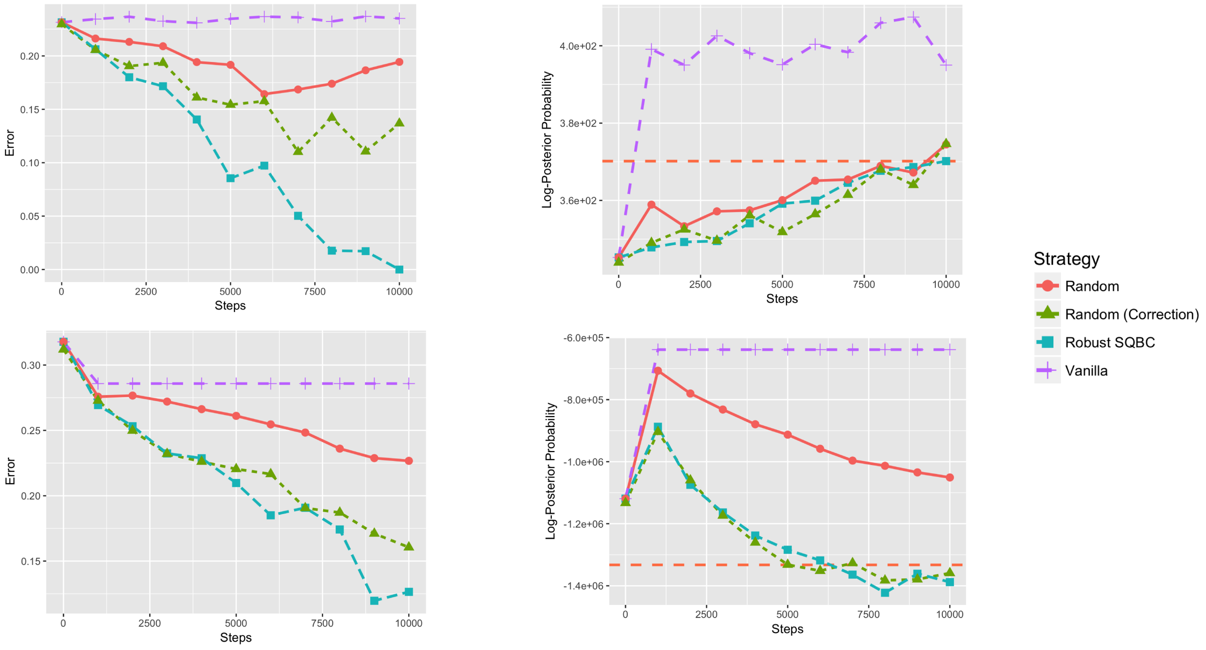


Figure 8.4. Mixture of Gaussians clustering experiments. x-axis corresponds to full passes of Gibbs sampler. Dashed orange line corresponds to log-posterior of target clustering. *Top*: `Iris` dataset. *Bottom*: `Wine` dataset.

Mixture of Bernoullis

Consider the following Bayesian generative model for a mixture of k Bernoulli product distributions:

- Weight vector $(w_1, \ldots, w_k)$ is drawn from a symmetric Dirichlet distribution with parameter $\alpha > 0$

- Bias variables $a_j^{(i)} \in [0, 1]$ are drawn i.i.d. from $\text{Beta}(\beta, \gamma)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, d$

- For each data point $i = 1, \ldots, n$:

  - $z_i \in \{1, \ldots, k\}$ is drawn from $\text{Categorical}(w_1, \ldots, w_k)$

  - $x_j^{(i)} \in \{0, 1\}$ is drawn from $\text{Bern}(a_j^{(z_i)})$ for $j = 1, \ldots, d$

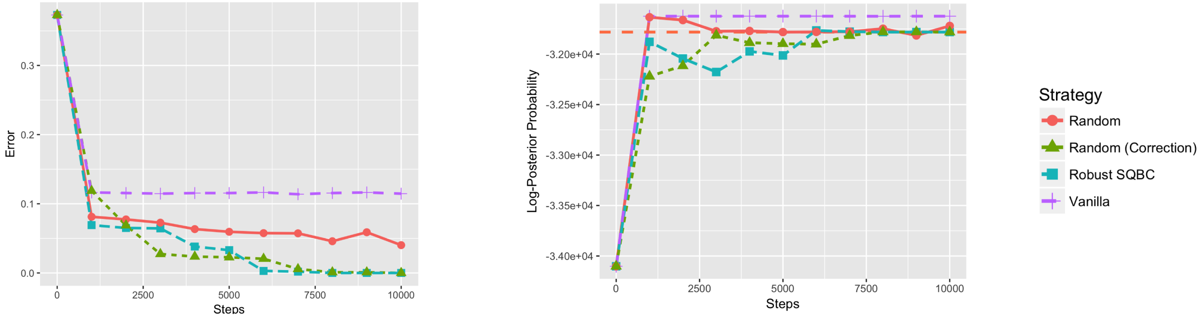


Figure 8.5. Mixture of Bernoullis experiments on MNIST dataset. Dashed orange line corresponds to log-posterior of target clustering.

Here, $x^{(1)}, \ldots, x^{(n)}$ are the observed d-dimensional binary-valued vectors, the a's are unobserved d-dimensional real-valued vectors, and α, β, and γ are known hyperparameters.

We ran experiments on a binarized version of the MNIST handwritten digit dataset [65]. Here we randomly sampled 50 images from each of the 0, 1, and 2 classes and sought to recover the clusters induced by these classes. The results are presented in Figure 8.5.

8.7.3 Linear separators

We also considered the classical active learning setting of linear separators. In all our experiments, we used QBC with a spherical prior distribution $N(0, I_d)$ and the squared-loss posterior update.

Noisy linear simulations

When learning a linear separator under classification noise, there is a true linear separator $h^* \in \mathbb{R}^d$. When a point $x \in \mathbb{R}^d$ is queried, we observe

$$y = \begin{cases} \operatorname{sign}(\langle h^*, x \rangle) & \text{with probability } 1 - p \\ -\operatorname{sign}(\langle h^*, x \rangle) & \text{with probability } p \end{cases}$$

In our simulations, we used various settings of both the noise level p and the

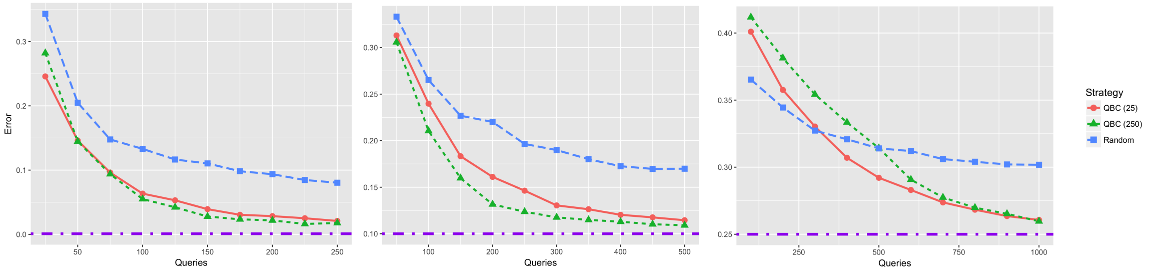


Figure 8.6. Simulations under different settings of the classification noise p. The dashed purple line is the level of classification noise. In the legend, QBC run with posterior update β is shown as 'QBC (β).'

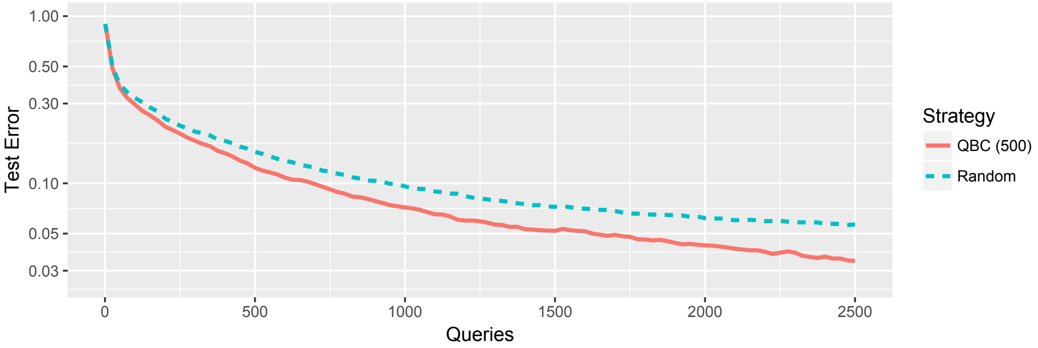


Figure 8.7. Kernel experiments on MNIST. As in Figure 8.6, QBC run with posterior update β is shown as 'QBC (β)' in the legend. Note that the test error axis is log-scale.

aggressiveness of the posterior update β. In the low noise setting, we found that setting β large appears to be appropriate. But as the noise level grows, using a smaller β appears to be advantageous. In all settings, however, QBC outperforms random sampling. Figure 8.6 shows some results of our simulations.

MNIST kernel experiments

We used the full MNIST handwritten digit dataset for our kernelized linear classification experiment. For this multiclass classification task, we interactively trained a collection of ten one-vs-all classifiers using the kernelized squared-loss approach outlined in Section 8.4. We used an RBF kernel $K(x, x') = \exp(-\gamma\|x - x'\|^2)$ with the choice of $\gamma = 0.001$. Because we are observing the true labels, we found that large values of β worked well for this task. Figure 8.7 shows the results in this setting. In particular, the experiments show the error obtained after 2500 random labels was obtained after just 1250 actively-sampled labels.

8.8 Related work

In this section, we discuss two areas of work related to the current manuscript: active learning and interactive clustering.

Active learning

For any overview of the active learning literature, see the related work section of Chapter 7.

The work presented here bears resemblance to both parametric [28, 34, 15] and nonparametric [90, 33, 29] active learning. On the one hand, our definition of structure is general enough to accommodate arbitrary labelings of a data set, as in nonparametric active learning. However, the algorithms considered here are much more closely related to those in the parametric setting. In particular, the algorithms considered in Section 8.3 are clear generalizations of the query-by-committee algorithm, while the robustified algorithm presented in Section 8.7 is in some sense an interpolation between the generalized binary search algorithm [30, 48, 76] and the DBAL algorithm of Chapter 7. However, where this work most clearly departs from the classical active learning literature is the change from question-answering to partial correction feedback.

The analysis we presented in this chapter is closely related to that of the original query-by-committee algorithm [41]; in particular, it uses a very similar notion of uncertainty. Our results are also very close in spirit to those obtained for generalized binary search [76]. In fact, the latter work is able to characterize the query complexity of binary active learning using certain geometric quantities, and it is an interesting open problem whether something similar can be done in our structural setting.

Interactive clustering

Another area relevant to this work is interactive clustering, which has also developed along several fronts.

One interactive clustering model considered in the literature allows users to split and merge clusters in the algorithm's current clustering until the target clustering has been found. In order to limit the amount of work required of the user, split requests are not allowed to specify how a cluster is split, only that a cluster should be split, and the algorithm must best decide how to go about splitting the chosen cluster. Under certain assumptions on the user's feedback and the target clustering, various algorithms have been shown to recover the target clustering while needing relatively few rounds of interaction with the user [14, 10, 8].

Another approach to interactive clustering is the *clustering with constraints* framework. In this model, a user provides constraints that the target clustering satisfies, and the algorithm attempts to find a clustering that both satisfies these constraints as well as optimizes some cost function. In the flat clustering setting, the constraints are very often must-link/cannot-link pairs [86]; while in the hierarchical clustering setting, there has been work on constraints that are ordered triplets indicating which points are closer in tree distance [85]. In the case where the target flat clustering optimizes the k-means cost function, it has been shown that relatively few queries can transform an NP-hard unsupervised learning problem into a tractable interactive learning problem [6].

As discussed in Section 8.2, the constraints in both the flat clustering and hierarchical clustering models can be written using in our interactive structure learning framework. It is less clear, however, how to incorporate split and merge feedback into our model. It thus remains an interesting open problem to marry fine-grained interaction such as must-link/cannot-link constraints with higher-level feedback such as cluster splits and merges.

Chapter 8 contains material that is currently being prepared for submission for publication of the material. C. Tosh and S. Dasgupta. The dissertation author was the primary investigator.

Appendix A

Supplementary material for Chapter 3

A.1 Proofs from Section 3.2

We start with a general result about the polynomial approximability of discrete distributions, and then consider an application to topic models.

Lemma A.1. *Consider any distribution with finite support, say $p = (p_1, \ldots, p_\ell)$. Pick any positive integer M. Then there is a distribution $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_\ell)$ such that:*

- *Each $\widehat{p}_i$ is a non-zero multiple of $1/M$.*
- *For each i, we have $\widehat{p}_i \geq (1 - \ell/M)p_i$.*

Proof. First define an intermediate distribution $\overline{p}$ as follows:

$$\overline{p}_i = (1 - \ell/M)p_i, \text{ rounded up to the nearest multiple of } 1/M.$$

Therefore, $(1 - \ell/M)p_i \leq \overline{p}_i \leq (1 - \ell/M)p_i + (1/M)$, and $\sum_i \overline{p}_i$ is some multiple of $1/M$ that is ≤ 1. To get $\widehat{p}$, take $\overline{p}$ and add multiples of $1/M$ to any coordinate(s) until the sum of the coordinates equals 1. □

This construction can be used to show that the maximum-likelihood topic model admits a concise approximation.

Lemma 3.2. *Consider any $V \times K$ topic distribution matrix Ψ. For any $\epsilon > 0$ and any integer m, there is a topic matrix $\widehat{\Psi}$ that uses $\lceil \log_2(mV/\epsilon) \rceil$ bits per entry, such that $\log p(x|\Psi) - \log p(x|\widehat{\Psi}) \leq \epsilon$ for all documents x of $\leq m$ words.*

Proof. Obtain $\widehat{\Psi}$ by applying the previous lemma to each individual topic distribution, with $M = \lceil 2mV/\epsilon \rceil$. Pick any document x of length m. Letting q denote the prior on topic weights (that is, a prior on the K-simplex), and letting $z \in \{1, \ldots, K\}^m$ denote the topic assignments to the words x_i, we have

$$\begin{aligned}\Pr(x|\Psi) &= \int q(\theta) \sum_z \Pr(z|\theta) \Pr(x|z, \Psi) \, d\theta \\ &= \sum_z \left(\prod_{i=1}^n \Psi_{x_i}^{(z_i)} \right) \int q(\theta) \Pr(z|\theta) \, d\theta\end{aligned}$$

By construction, for any z,

$$\prod_{i=1}^m \widehat{\Psi}_{x_i}^{(z_i)} \;\geq\; \prod_{i=1}^m \left((1 - \epsilon/2m) \Psi_{x_i}^{(z_i)} \right) \;=\; (1 - \epsilon/2m)^m \prod_{i=1}^m \Psi_{x_i}^{(z_i)} \;\geq\; e^{-\epsilon} \prod_{i=1}^m \Psi_{x_i}^{(z_i)},$$

and thus $\Pr(x|\Psi) \leq e^{\epsilon} \Pr(x|\widehat{\Psi})$, as claimed. □

A.1.1 Proof of Theorem 3.3: hardness of finding the maximum-likelihood topic model

Our goal is to prove the following theorem.

Theorem 3.3. *[Implicit in [4]] We say a topic matrix Ψ is* c-smooth *for $c > 0$ if $\min_i \max_j \Psi_i^{(j)} \geq c$. Given $\alpha > 0$,* TM-MLE(α) *is NP-hard when $K = 2$, all the documents are restricted to have 2 words, and Ψ_{ML} is guaranteed to be $(1/V)$-smooth.*

In fact, we will prove a more general result. Let Δ^N be the N-simplex, i.e.

$$\Delta^N \;=\; \left\{ \theta \in \mathbb{R}^N : \sum_{i=1}^N \theta_i = 1 \text{ and } \theta_i \geq 0 \right\}.$$

Theorem A.2. *Let $\lambda_S, \lambda_X \geq 0$ and ν_0 be a distribution over Δ^K such that for $\theta \sim \nu_0$*

- $\mathbb{E}[\theta_1^2] = \cdots = \mathbb{E}[\theta_k^2] = \lambda_S$ *and*
- $\mathbb{E}[\theta_i \theta_j] = \lambda_X$ *for all $i \neq j$.*

Then TM-MLE-ν_0*, the problem of maximizing the same objective as* TM-MLE(α) *with the Dir(α) prior replaced with ν_0, is NP-hard when $\lambda_S > \lambda_X$ when $K = 2$, there are exactly two words in each document, and the ML solution is $1/V$-smooth.*

To see how this implies Theorem 3.3, note that for $\theta \sim \mathrm{Dir}(\alpha)$ and $i \neq j$,

$$\begin{aligned}\lambda_S = \mathbb{E}[\theta_i^2] = \frac{\Gamma(K\alpha)\Gamma(\alpha+2)}{\Gamma(K\alpha+2)\Gamma(\alpha)} = \frac{\alpha+1}{K(\alpha K+1)} &> \frac{\alpha}{K(\alpha K+1)} \\ &= \frac{\Gamma(K\alpha)(\Gamma(\alpha+1))^2}{\Gamma(K\alpha+2)(\Gamma(\alpha))^2} = \mathbb{E}[\theta_i\theta_j] = \lambda_X.\end{aligned}$$

The proof follows the reduction from Arora et al. [4] very closely. We start from an instance of MINIMUM-BISECTION. Here the input is a graph $G = (V, E)$ with $|V| = n$ even and $|E| = m$, and the goal is to find a cut (S, T) such that $|S| = n/2 = |T|$ and $|E(S,T)|$ is minimized.

Beginning with G, we construct our instance of TM-MLE-ν_0 as follows. The vocabulary is the set of vertices V. Our corpus will consist of the following documents:

- for each word $i \in V$, create N documents with only the word i repeated twice, and
- for each edge $(i, j) \in E$, create one document with only the word i and the word j.

Here N is a polynomial of n, m, λ_S, and λ_X to be determined later. Given a document with words i and j (possibly equal) and a topic matrix $\Psi = [\Psi^{(1)}|\Psi^{(2)}]$, what is the likelihood of the document under Ψ? This is simply

$$p(i, j \mid \Psi) = \mathbb{E}[\theta_1^2]\Psi_i^{(1)}\Psi_j^{(1)} + \mathbb{E}[\theta_1\theta_2]\left(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}\right) + \mathbb{E}[\theta_2^2]\Psi_i^{(2)}\Psi_j^{(2)}$$

$$= \lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X \left(\Psi_i^{(1)} \Psi_j^{(2)} + \Psi_j^{(1)} \Psi_i^{(2)} \right)$$

where $\Psi_i = (\Psi_i^{(1)}, \Psi_i^{(2)})$. Then the objective is to maximize the following function:

$$\begin{aligned}
F(\Psi) &= \sum_{\text{document}=(i,j)} \ln \left(\lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X (\Psi_i^{(1)} \Psi_j^{(2)} + \Psi_j^{(1)} \Psi_i^{(2)}) \right) \\
&= \sum_{i \in V} N \ln \left(\lambda_S \|\Psi_i\|_2^2 + 2\lambda_X \Psi_i^{(1)} \Psi_i^{(2)} \right) \\
&\qquad + \sum_{(i,j) \in E} \ln \left(\lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X (\Psi_i^{(1)} \Psi_j^{(2)} + \Psi_j^{(1)} \Psi_i^{(2)}) \right).
\end{aligned}$$

For any bisection (S, T), define the *canonical solution* $\Psi = \Psi(S, T)$ to be the topic matrix which satisfies $\Psi_i^{(1)} = 2/n$ and $\Psi_i^{(2)} = 0$ for all $i \in S$; and $\Psi_i^{(1)} = 0$ and $\Psi_i^{(2)} = 2/n$ for all $i \in T$. We'll see that the maximum-likelihood solution (or an approximation thereof) is approximately canonical, and therefore uniquely specifies a cut.

Write $F(\Psi) = G(\Psi) + H(\Psi)$, where

$$\begin{aligned}
G(\Psi) &= \sum_{i \in V} N \ln \left(\lambda_S \|\Psi_i\|_2^2 + 2\lambda_X \Psi_i^{(1)} \Psi_i^{(2)} \right) \\
H(\Psi) &= \sum_{(i,j) \in E} \ln \left(\lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X (\Psi_i^{(1)} \Psi_j^{(2)} + \Psi_j^{(1)} \Psi_i^{(2)}) \right).
\end{aligned}$$

When N is made large enough, G dominates H.

Each of the n rows of Ψ is a pair $(\Psi_i^{(1)}, \Psi_i^{(2)})$. We start by characterizing approximately optimal solutions subject to specific row-sum constraints.

Lemma A.3. *Suppose the row-sums are constrained to be $\Psi_i^{(1)} + \Psi_i^{(2)} = r_i$, for some $r_1, \ldots, r_n$ summing to 2. Then:*

(a) G is bounded as follows:

$$G(\Psi) \;\le\; \sum_{i=1}^{n} N \ln(\lambda_S r_i^2) - N \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}$$

with equality if each row has $\min(\Psi_i^{(1)}, \Psi_i^{(2)}) = 0$.

(b) H lies in a smaller range:

$$m \ln \lambda_X \;\leq\; H(\Psi) - \sum_{(i,j)\in E} \ln(r_i r_j) \;\leq\; m \ln \lambda_S.$$

Proof. To see (a), first note that

$$\frac{2\Psi_i^{(1)}\Psi_i^{(2)}}{r_i^2} = \frac{2\Psi_i^{(1)}\Psi_i^{(2)}}{(\Psi_i^{(1)}+\Psi_i^{(2)})^2} \geq \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{\Psi_i^{(1)}+\Psi_i^{(2)}} = \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}.$$

Therefore, we can write

$$\begin{aligned}
G(\Psi) &= \sum_{i=1}^{n} N \ln\left(\lambda_S \|\Psi_i\|_2^2 + 2\lambda_X \Psi_i^{(1)}\Psi_i^{(2)}\right) \\
&= \sum_{i=1}^{n} N \ln\left(\lambda_S r_i^2\right) + N \sum_{i=1}^{n} \ln\left(\frac{r_i^2 - 2\Psi_i^{(1)}\Psi_i^{(2)}}{r_i^2} + \frac{2\lambda_X \Psi_i^{(1)}\Psi_i^{(2)}}{\lambda_S r_i^2}\right) \\
&= \sum_{i=1}^{n} N \ln\left(\lambda_S r_i^2\right) + N \sum_{i=1}^{n} \ln\left(1 - \frac{2\Psi_i^{(1)}\Psi_i^{(2)}}{r_i^2} \cdot \frac{\lambda_S - \lambda_X}{\lambda_S}\right) \\
&\leq \sum_{i=1}^{n} N \ln\left(\lambda_S r_i^2\right) - N \sum_{i=1}^{n} \frac{2\Psi_i^{(1)}\Psi_i^{(2)}}{r_i^2} \cdot \frac{\lambda_S - \lambda_X}{\lambda_S} \\
&\leq \sum_{i=1}^{n} N \ln(\lambda_S r_i^2) - N \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}.
\end{aligned}$$

(b) follows directly from algebra and applying the inequality $\lambda_S > \lambda_X$. □

This immediately gives us the following corollary.

Corollary A.4. *Fix any row-sums $r_1, \ldots, r_n$, and any $\Delta > 0$. Let Ψ be any solution whose $F(\cdot)$ value is within Δ of optimal, subject to these constraints. Then for each i,*

$$\frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i} \;\leq\; \frac{1}{N}\left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta\lambda_S}{\lambda_S - \lambda_X}\right).$$

Thus each row has one entry that is approximately zero, whereupon, returning to Lemma A.3, we see that $G(\cdot)$ is roughly $2N \sum_i \ln r_i$, ignoring constants. The following technical lemma then implies that in an approximately optimal solution, all row-sums must be roughly equal.

Lemma A.5. *Subject to the constraint that $r_1, \ldots, r_n$ are nonnegative and sum to 2:*

(a) The quantity $\sum_i \ln r_i$ is maximized when the r_i are equal, in which case

$$\sum_{i=1}^{n} \ln r_i \;=\; n \ln \frac{2}{n}.$$

(b) Pick any $\epsilon > 0$. If there is some $r_i \notin [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$, then no matter how the other r_j are set,

$$\sum_{i=1}^{n} \ln r_i \;\leq\; n \ln \frac{2}{n} - \frac{1}{4}\epsilon^2.$$

Proof. (a) follows directly from Jensen's inequality. To see (b), we make use of the following logarithmic inequalities, which can be found in Topsøe [84]. For $0 \leq x < 1$,

$$\ln(1+x) \leq \frac{x}{2} \cdot \frac{2+x}{1+x} \qquad \text{and} \qquad \ln(1-x) \leq \frac{-2x}{2-x}.$$

Now let $\delta > 0$ and suppose that there is some i such that $r_i = 2(1+\delta)/n$. Then by (a),

$$\begin{aligned}
\sum_{j=1}^{n} \ln r_j &= \ln\left(\frac{2}{n}(1+\delta)\right) + \sum_{j \neq i} \ln r_j \\
&\leq \ln\left(\frac{2}{n}(1+\delta)\right) + (n-1)\ln\left(\frac{1}{n-1}\left(2 - \frac{2}{n}(1+\delta)\right)\right) \\
&= n \ln \frac{2}{n} + \ln(1+\delta) + (n-1)\ln\left(1 - \frac{\delta}{n-1}\right) \\
&\leq n \ln \frac{2}{n} + \frac{\delta}{2} \cdot \frac{2+\delta}{1+\delta} - \frac{2\delta(n-1)}{2(n-1)-\delta} \\
&\leq n \ln \frac{2}{n} - \frac{1}{4}\delta^2
\end{aligned}$$

They proof for the case where $r_i = 2(1-\delta)/n$ is similar. Thus, we have (b). □

The $G(\cdot)$ function dominates $H(\cdot)$ and forces (approximately) canonical solutions.

Lemma A.6. *Pick any $0 < \epsilon < 1$ and any $\Delta > 0$. Define*

$$\begin{aligned} N_0 &= \frac{2}{\epsilon^2}\left(\Delta + m \ln\left(\frac{\lambda_S}{\lambda_X}\frac{n^2}{4}\right)\right) \\ N_1 &= \frac{1}{\epsilon}\left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta\lambda_S}{\lambda_S - \lambda_X}\right). \end{aligned}$$

Let Ψ^* be a maximizer of $F(\cdot)$. Then, if $N \geq \max(N_0, N_1)$, any solution Ψ with $F(\Psi) \geq F(\Psi^*) - \Delta$ must satisfy the following conditions for each i:

(a) $\Psi_i^{(1)} + \Psi_i^{(2)} \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$.

(b) $\min(\Psi_i^{(1)}, \Psi_i^{(2)}) \leq \epsilon \cdot \frac{2}{n}$.

Proof. Let Φ be any canonical solution (which trivially implies $F(\Psi^*) \geq F(\Phi)$), let Ψ be a solution satisfying $F(\Psi) \geq F(\Psi^*) - \Delta$, and let $r_1, \ldots, r_n$ be the row sums of Ψ. Then because Φ is canonical, we know from the above lemmas

$$\begin{aligned} \Delta &\geq F(\Phi) - F(\Psi) \\ &= G(\Phi) - G(\Psi) + H(\Phi) - H(\Psi) \\ &\geq G(\Phi) - G(\Psi) + \sum_{(i,j)\in E} \ln\left(\frac{4}{n^2}\right) + m \ln \lambda_X - \sum_{(i,j)\in E} \ln(r_i r_j) - m \ln \lambda_S \\ &= G(\Phi) - G(\Psi) - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) - \sum_{(i,j)\in E} \ln(r_i r_j) \\ &\geq N\left(n \ln\left(\lambda_S \frac{4}{n^2}\right) - \sum_{i=1}^{n} \ln(\lambda_S r_i^2) + \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i}\right) \\ &\quad - m \ln\left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X}\right) - \sum_{(i,j)\in E} \ln(r_i r_j). \end{aligned}$$

Now suppose by contradiction that Ψ does not satisfy condition (a). We know that because the columns of Ψ sum to 1, it must be the case that $r_i r_j \leq 1$. Applying Lemma A.5(b),

$$\begin{aligned}
\Delta &\geq N \left(n \ln \left(\lambda_S \frac{4}{n^2} \right) - \sum_{i=1}^{n} \ln(\lambda_S r_i^2) + \sum_{i=1}^{n} \frac{\lambda_S - \lambda_X}{\lambda_S} \frac{\min(\Psi_i^{(1)}, \Psi_i^{(2)})}{r_i} \right) - m \ln \left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X} \right) \\
&> N \left(n \ln \left(\lambda_S \frac{4}{n^2} \right) - n \ln \left(\lambda_S \frac{4}{n^2} \right) + \frac{\epsilon^2}{2} \right) - m \ln \left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X} \right) \\
&= \frac{N\epsilon^2}{2} - m \ln \left(\frac{n^2}{4} \cdot \frac{\lambda_S}{\lambda_X} \right).
\end{aligned}$$

But this implies that

$$N < \frac{2}{\epsilon^2} \left(\Delta + m \ln \left(\frac{\lambda_S}{\lambda_X} \frac{n^2}{4} \right) \right) = N_0$$

which is a contradiction.

To see that Ψ must satisfy condition (b), note that by Corollary A.4, if Ψ did not satisfy condition (b), then $F(\Psi)$ could not be within Δ of $F(\Psi^*)$. □

Once we are within the realm of approximately canonical solutions, which uniquely designate a bisection cut, the lower-order term $H(\cdot)$ serves to choose a cut of small size.

Lemma A.7. *Pick any $0 < \epsilon < 1$. We will describe any Ψ that satisfies conditions (a) and (b) of Lemma A.6 as being ϵ-approximately canonical.*

(a) For any canonical solution Ψ,

$$H(\Psi) \;=\; m \ln \frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X}.$$

(b) For any ϵ-approximately canonical solution Ψ,

$$H(\Psi) \;\leq\; m \ln \frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X} + 2m\epsilon \frac{\lambda_S}{\lambda_X}.$$

Proof. Recall that

$$H(\Psi) = \sum_{(i,j)\in E} \ln\left(\lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})\right).$$

Therefore, if Ψ is a canonical solution corresponding to the bisection (S,T), then if $E(S,T)$ denotes the subset of edges with one endpoint in S and the other in T we have

$$\begin{aligned} H(\Psi) &= \sum_{(i,j)\in E(S,T)} \ln\frac{4\lambda_X}{n^2} + \sum_{(i,j)\in E\setminus E(S,T)} \ln\frac{4\lambda_S}{n^2} \\ &= m\ln\frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi)|\cdot\ln\frac{\lambda_S}{\lambda_X}. \end{aligned}$$

Now let Ψ be an ϵ-approximately canonical solution. Use it to define a cut (S,T) in the natural way:

$$S = \{i : \Psi_i^{(2)} \le 2\epsilon/n\}, \quad T = [n]\setminus S.$$

Given an edge $(i,j)\in E$, how do we bound $Q_{i,j}(\Psi) = \lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)})$? We consider two cases.

Case 1: $(i,j)\in E\setminus E(S,T)$. Assume w.l.o.g. that $i,j\in S$. Then because Ψ is ϵ-approximately canonical, we know $\|\Psi_i\|_1, \|\Psi_j\|_1 \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$ and $\Psi_i^{(2)}, \Psi_j^{(2)} \le \frac{2}{n}\epsilon$. Letting $\Psi_i^{(2)} = \frac{2}{n}\delta_i$ and $\Psi_j^{(2)} = \frac{2}{n}\delta_j$, we have

$$\begin{aligned} Q_{i,j}(\Psi) &= \lambda_S\langle\Psi_i,\Psi_j\rangle + \lambda_X(\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}) \\ &\le \frac{4}{n^2}\left(\lambda_S((1+\epsilon-\delta_i)(1+\epsilon-\delta_j) + \delta_i\delta_j) + \lambda_X((1+\epsilon-\delta_j)\delta_i + (1+\epsilon-\delta_i)\delta_j)\right) \\ &= \frac{4}{n^2}\left(\lambda_S(1+\epsilon)^2 + (\lambda_S-\lambda_X)(2\delta_i\delta_j - (1+\epsilon)(\delta_i+\delta_j))\right) \end{aligned}$$

Since $\delta_i, \delta_j \le \epsilon < 1$ and $\lambda_S > \lambda_X$, the above is maximized whenever $\delta_i = \delta_j = 0$. Thus,

$$Q_{i,j}(\Psi) \le \frac{4\lambda_S(1+\epsilon)^2}{n^2}.$$

Case 2: $(i, j) \in E(S, T)$. Assume w.l.o.g. that $i \in S$ and $j \in T$. Then because Ψ is ϵ-approximately canonical, we know $\|\Psi_i\|_1, \|\Psi_j\|_1 \in [\frac{2}{n}(1-\epsilon), \frac{2}{n}(1+\epsilon)]$ and $\Psi_i^{(2)}, \Psi_j^{(1)} \leq \frac{2}{n}\epsilon$. Letting $\Psi_i^{(2)} = \frac{2}{n}\delta_i$ and $\Psi_j^{(1)} = \frac{2}{n}\delta_j$, we have

$$\begin{aligned}
Q_{i,j}(\Psi) &= \lambda_S \langle \Psi_i, \Psi_j \rangle + \lambda_X (\Psi_i^{(1)}\Psi_j^{(2)} + \Psi_j^{(1)}\Psi_i^{(2)}) \\
&\leq \frac{4}{n^2} \left(\lambda_S((1+\epsilon)\delta_i + (1+\epsilon)\delta_j) + \lambda_X((1+\epsilon-\delta_i)(1+\epsilon-\delta_j) + \delta_i\delta_j \right) \\
&= \frac{4}{n^2} \big(\lambda_X(1+\epsilon)^2 + (\lambda_S - \lambda_X)((1+\epsilon)(\delta_i+\delta_j) - \delta_i\delta_j \big) \\
&\leq \frac{4}{n^2} \big(\lambda_X(1+\epsilon)^2 + 2(\lambda_S - \lambda_X)(1+\epsilon)\epsilon \big) \\
&= \frac{4\lambda_X}{n^2}(1+\epsilon)\left(1 + \frac{\epsilon(2\lambda_S - \lambda_X)}{\lambda_X}\right)
\end{aligned}$$

Combining the above two cases, we can bound on $H(\Psi)$ above by

$$\begin{aligned}
&\sum_{(i,j)\in E\setminus E(S,T)} \ln\left(\frac{4\lambda_S(1+\epsilon)^2}{n^2}\right) + \sum_{(i,j)\in E(S,T)} \ln\left(\frac{4\lambda_X}{n^2}(1+\epsilon)\left(1+\frac{\epsilon(2\lambda_S-\lambda_X)}{\lambda_X}\right)\right) \\
&= m \ln\frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi)| \cdot \ln\frac{\lambda_S}{\lambda_X} + |\mathrm{cut}(\Psi)| \ln\left((1+\epsilon)\left(1+\frac{\epsilon(2\lambda_S-\lambda_X)}{\lambda_X}\right)\right) \\
&\qquad + (m - |\mathrm{cut}(\Psi)|)\ln((1+\epsilon)^2) \\
&\leq m \ln\frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi)| \cdot \ln\frac{\lambda_S}{\lambda_X} + m\epsilon \max\left(2, 1+\frac{2\lambda_S-\lambda_X}{\lambda_X}\right).
\end{aligned}$$

Using the fact that $\lambda_S > \lambda_X$ gives us the lemma. □

Let $\Delta, \epsilon, N_0, N_1, N > 0$ satisfy the relationship specified in Lemma A.6. We will argue that for an appropriate, but polynomial setting, of these variables, any Δ-optimal solution must correspond to the minimum bisection.

Let Ψ be a Δ-optimal solution. By Lemma A.6, Ψ must be ϵ-approximately canonical. As in the proof of Lemma A.7, we can use Ψ to define a cut (S, T). For $\epsilon < 1/(2n)$, this cut is a bisection. Now let (S^*, T^*) be an optimal bisection and let Ψ^* be the solution corresponding to this. Then we can say

$$\begin{aligned}\Delta &\geq \max_{\Psi'} F(\Psi') - F(\Psi) \geq F(\Psi^*) - F(\Psi) \\ &= G(\Psi^*) - G(\Psi) + H(\Psi^*) - H(\Psi) \geq H(\Psi^*) - H(\Psi).\end{aligned}$$

Now by Lemma A.7, we have

$$\begin{aligned}\Delta &\geq H(\Psi^*) - H(\Psi) \\ &\geq m \ln \frac{4\lambda_S}{n^2} - |\mathrm{cut}(\Psi^*)| \cdot \ln \frac{\lambda_S}{\lambda_X} - m \ln \frac{4\lambda_S}{n^2} + |\mathrm{cut}(\Psi)| \cdot \ln \frac{\lambda_S}{\lambda_X} - 2m\epsilon \frac{\lambda_S}{\lambda_X} \\ &= (|\mathrm{cut}(\Psi)| - |\mathrm{cut}(\Psi^*)|) \ln\left(\frac{\lambda_S}{\lambda_X}\right) - 2m\epsilon \frac{\lambda_S}{\lambda_X} \\ &\geq (|\mathrm{cut}(\Psi)| - |\mathrm{cut}(\Psi^*)|) \left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right) - 2m\epsilon \frac{\lambda_S}{\lambda_X}.\end{aligned}$$

Thus, if $\Delta \leq \frac{1}{3}\left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right)$ and $\epsilon \leq \frac{1}{6m}\left(\frac{\lambda_X}{\lambda_S}\right)\left(\frac{\lambda_S - \lambda_X}{\lambda_S}\right)$, then we must conclude that

$$|\mathrm{cut}(\Psi)| = |\mathrm{cut}(\Psi^*)|.$$

These settings of ϵ and Δ, give us

$$\begin{aligned}N_0 &= \frac{2}{\epsilon^2}\left(\Delta + m \ln\left(\frac{\lambda_S}{\lambda_X}\frac{n^2}{4}\right)\right) \\ &= 72m^2 \left(\frac{\lambda_S}{\lambda_X}\right)^2 \left(\frac{\lambda_S}{\lambda_S - \lambda_X}\right)^2 \left((m + 1/2) \ln\left(\frac{\lambda_S}{\lambda_X}\right) + m \ln\left(\frac{n^2}{4}\right)\right)\end{aligned}$$

and

$$\begin{aligned}N_1 &= \frac{1}{\epsilon}\left(\frac{m\lambda_S}{\lambda_X} + \frac{\Delta\lambda_S}{\lambda_S - \lambda_X}\right) \\ &= 6m \left(\frac{\lambda_S}{\lambda_X}\right)\left(\frac{\lambda_S}{\lambda_S - \lambda_X}\right)\left(\frac{m\lambda_S}{\lambda_X} + \frac{\lambda_S}{2(\lambda_S - \lambda_X)} \ln\left(\frac{\lambda_S}{\lambda_X}\right)\right).\end{aligned}$$

This completes the proof of Theorem 3.3.

A.2 Proofs from Section 3.3

A.2.1 Proof of Lemma 3.6

The goal of this section is to prove the following lemma.

Lemma 3.6. *Let $c, m > 0$ and suppose that Ψ and Φ are $V \times K$ c-smooth topic matrices such that $\|\Psi - \Phi\|_{max} \leq \min(c/m, 1/2)$. If $\alpha_0 = \sum \alpha_i$, then for any document x with length bounded by m,*

$$|\log p(x \mid \Psi) - \log p(x \mid \Phi)| \;\leq\; \|\Psi - \Phi\|_{max} \left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m} \right).$$

To do so, we need to introduce some notation. Suppose $x = (i_1, \ldots, i_m)$ is some length m document. Then $z \in [K]^m$ is a *labeling* of x, that is an assignment of each word in x to some topic. For some fixed labeling z, let $n_i(z) = \{j \,:\, z_j = i\}$ denote the number of times that Define the likelihood of z under Ψ is given by

$$q(\Psi, z) = \left(\prod_{i=1}^{K} \frac{\Gamma(\alpha_i + n_i(z))}{\Gamma(\alpha_i)} \right) \prod_{j=1}^{m} \Psi_{i_j}^{(z_j)}.$$

Then we see that summing over all labelings gives us the log-likelihood of document x.

Lemma A.8. *For any length m document x and any topic matrix Ψ,*

$$p(x|\Psi) = \sum_{z \in [K]^m} q(\Psi, z).$$

Proof. To generate document $x = (i_1, \ldots, i_m)$ given Ψ, we can first sample $\theta \sim Dir(\alpha_1, \ldots, \alpha_K)$. Given θ, we can sample $z_1, \ldots, z_m$ independently from the distribution θ and then independently sample each word j from the distribution Ψ^{z_j}. Marginalizing over θ and z and recognizing that x is independent of θ given the z's,

$$\begin{aligned} p(x \mid \Psi) &= \mathbb{E}_\theta[p(x \mid \Psi, \theta)] \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[p(z|\theta) p(x|\Psi, \theta, z) \right] \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[\prod_{j=1}^{m} \theta_{z_j} \right] \prod_{j=1}^{m} \Psi_{i_j}^{z_j} \\ &= \sum_{z \in [K]^m} \mathbb{E}_\theta \left[\prod_{i=1}^{K} \theta_i^{n_i(z)} \right] \prod_{j=1}^{m} \Psi_{i_j}^{z_j} \end{aligned}$$

The expectation in the last line deals with the moments of the Dirichlet distribution. [75] provides the following identity for the moments of the Dirichlet distribution

$$\mathbb{E}_\theta \left[\prod_{i=1}^{k} \theta_i^{n_i} \right] = \frac{\Gamma(\sum \alpha_i)}{\Gamma(\sum \alpha_i + n_i)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}$$

for positive integers $n_1, \ldots, n_K$. Plugging this into the above gives us the lemma. □

Therefore, proving Lemma 3.6 amounts to getting a handle on a particular ratio of sums:

$$\frac{p(x \mid \Psi)}{p(x \mid \Phi)} = \frac{\sum_{z \in [K]^n} q(\Psi, z)}{\sum_{z \in [K]^n} q(\Phi, z)}.$$

The next few technical lemmas deal with bounding ratios of sums.

Ratios of sums

Lemma A.9. *Let $a_1, \ldots, a_n, b_1, \ldots, b_n, c > 0$ such that $a_i/b_i \leq c$, then $\frac{\sum a_i}{\sum b_i} \leq c$.*

Proof. We have that $a_i \leq cb_i$ for all i. Thus, $\frac{\sum a_i}{\sum b_i} \leq \frac{\sum cb_i}{\sum b_i} \leq c$. □

Lemma A.10. *Suppose $a, b, c, d, \epsilon > 0$, $x, y \in [0, 1]$, and $|x - y| \leq \epsilon$ then*

$$\frac{a + cx}{b + dy} \leq \max \left(\frac{a + c\epsilon}{b}, \frac{a + c}{b + d(1 - \epsilon)} \right).$$

Proof. There are two cases.

Case 1: $y \geq 1 - \epsilon$. In this case we have

$$\frac{a + cx}{b + dy} \leq \frac{a + c}{b + d(1 - \epsilon)}.$$

Case 2: $y \leq 1 - \epsilon$. In this case we have

$$\frac{a + cx}{b + dy} \leq \frac{a + c(y + \epsilon)}{b + dy} =: f(y)$$

Then it can be shown that the sign of f' is independent of y (since $y \geq 0$). Therefore f is monotonic in y and reaches the maximum at the boundary $\{0, 1 - \epsilon\}$. □

Lemma A.11. *Let $a, b, c, \epsilon > 0$ and $x_1, \ldots, x_n, y_1, \ldots, y_n \in [0, 1]$ such that $|x_i - y_i| < \epsilon$, then*

$$\frac{a + c\prod_{i=1}^{n} x_i}{b + c\prod_{i=1}^{n} y_i} \leq \max\left(\frac{a + \epsilon c}{b}, \frac{a + c}{b + (1 - \epsilon)^n c}\right).$$

Proof. The proof is by induction on n. The base case is simply an appeal to Lemma A.10. Now assume it holds for $n - 1$. There are three cases we need to consider.

Case 1: $y_n = 0$. In this case we know $x_n \leq \epsilon$, therefore

$$\frac{a + c\prod_{i=1}^{n} x_i}{b + c\prod_{i=1}^{n} y_i} \leq \frac{a + \epsilon c\prod_{i=1}^{n-1} x_i}{b} \leq \frac{a + \epsilon c}{b}.$$

Case 2: $x_n = 0$. In this case,

$$\frac{a + c\prod_{i=1}^{n} x_i}{b + c\prod_{i=1}^{n} y_i} \leq \frac{a}{b} \leq \frac{a + \epsilon c}{b}.$$

Case 3: $x_n, y_n > 0$. In this case we can use our inductive assumption to see the following

$$\begin{aligned}
\frac{a + c\prod_{i=1}^{n} x_i}{b + c\prod_{i=1}^{n} y_i} &= \frac{x_n}{y_n} \cdot \frac{a/x_n + c\prod_{i=1}^{n-1} x_i}{b/y_n + c\prod_{i=1}^{n-1} y_i} \\
&\leq \frac{x_n}{y_n} \max\left(\frac{a/x_n + \epsilon c}{b/y_n}, \frac{a/x_n + c}{b/y_n + (1-\epsilon)^{n-1}c}\right) \\
&= \max\left(\frac{a + \epsilon x_n c}{b}, \frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c}\right) \\
&\leq \max\left(\frac{a + \epsilon c}{b}, \frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c}\right).
\end{aligned}$$

By appealing again to Lemma A.10, we have

$$\frac{a + x_n c}{b + y_n(1-\epsilon)^{n-1}c} \leq \max\left(\frac{a + \epsilon c}{b}, \frac{a + c}{b + (1-\epsilon)^{n}c}\right).$$

Combining all of the above gives us the lemma. □

Lemma A.12. *Let $a, b, c_i, \epsilon > 0$ and $x_{i,j}, y_{i,j} \in [0,1]$ such that $|x_{i,j} - y_{i,j}|$ for $i \in [m], j \in [n]$, then there exists a partition Ω_1, Ω_2 of $[m]$*

$$\frac{a + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} x_{i,j}}{b + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} y_{i,j}} \leq \frac{a + \sum_{i\in\Omega_1} \epsilon c_i + \sum_{i\in\Omega_2} c_i}{b + \sum_{i\in\Omega_2} (1-\epsilon)^n c_i}.$$

Proof. We prove by induction on m. The base case of $m = 1$ follows directly from Lemma A.11. We can assume that the lemma holds for $m - 1$, then

$$\frac{a + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} x_{i,j}}{b + \sum_{i=1}^{m} c_i \prod_{j=1}^{n} y_{i,j}} = \frac{\overbrace{a + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}^{a'} + c_m \prod_{j=1}^{n} x_{m,j}}{\underbrace{b + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}}_{b'} + c_m \prod_{j=1}^{n} y_{m,j}}.$$

By applying Lemma A.11, we have that this is bounded by

$$\max\left(\frac{a' + \epsilon c_m}{b'}, \frac{a' + c_m}{b' + (1-\epsilon)^n c_m}\right).$$

We will bound each of these quantities separately. Denoting $a_1 = a + \epsilon c_m$, then by induction we have that there exists a partition Ω_1', Ω_2' of $[m-1]$ such that

$$\begin{aligned}\frac{a' + \epsilon c_m}{b'} &= \frac{a_1 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}{b + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}} \\ &\leq \frac{a_1 + \sum_{i\in\Omega_1'} \epsilon c_i + \sum_{i\in\Omega_2'} c_i}{b + \sum_{i\in\Omega_2'} (1-\epsilon)^n c_i} \\ &= \frac{a + \sum_{i\in\Omega_1'\cup\{m\}} \epsilon c_i + \sum_{i\in\Omega_2'} c_i}{b + \sum_{i\in\Omega_2'} (1-\epsilon)^n c_i}.\end{aligned}$$

On the other hand, if we let $a_2 = a + c_m$ and $b_2 = b + (1-\epsilon)^n c_m$, then by induction there exists a partition Ω_1'', Ω_2'' of $[m-1]$ such that

$$\begin{aligned}\frac{a' + c_m}{b' + (1-\epsilon)^n c_m} &= \frac{a_2 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} x_{i,j}}{b_2 + \sum_{i=1}^{m-1} c_i \prod_{j=1}^{n} y_{i,j}} \\ &\leq \frac{a_2 + \sum_{i\in\Omega_1''} \epsilon c_i + \sum_{i\in\Omega_2''} c_i}{b_2 + \sum_{i\in\Omega_2''} (1-\epsilon)^n c_i} \\ &= \frac{a + \sum_{i\in\Omega_1''} \epsilon c_i + \sum_{i\in\Omega_2''\cup\{m\}} c_i}{b + \sum_{i\in\Omega_2''\cup\{m\}} (1-\epsilon)^n c_i}.\end{aligned}$$

By taking Ω_1, Ω_2 to be the partitions corresponding to the larger of these two scenarios (either $\Omega_1' \cup \{m\}, \Omega_2'$ or $\Omega_1'', \Omega_2'' \cup \{m\}$), we have the lemma statement. □

Actual proof of Lemma 3.6

We are now ready to prove the main lemma of this section.

Lemma 3.6. *Let $c, m > 0$ and suppose that Ψ and Φ are $V \times K$ c-smooth topic matrices such that $\|\Psi - \Phi\|_{max} \leq \min(c/m, 1/2)$. If $\alpha_0 = \sum \alpha_i$, then for any document x with* *length bounded by* m*,*

$$|\log p(x\,|\,\Psi) - \log p(x\,|\,\Phi)| \;\leq\; \|\Psi - \Phi\|_{max}\left(\frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right)\frac{K^{\alpha_0+2m-1/2}}{c^m}\right).$$

Proof. Let $\Omega = [K]^m$ denote the space of labelings and let $x = (i_1, \ldots, i_m)$. From the smoothness condition on Ψ and Φ, we can see that there is a labeling $z^* \in \Omega$ such that $\Phi_{i_j}^{(z_j^*)} \geq c$ for $j = 1, \ldots, m$. From Lemma A.12 we know that we can partition $\Omega \setminus \{z^*\}$ into Ω_1, Ω_2 such that

$$\begin{aligned}
\frac{p(x\,|\,\Psi)}{p(x\,|\,\Phi)} &= \frac{q(\Psi, z^*) + \sum_{z\in\Omega\setminus\{z^*\}} q(\Psi, z)}{q(\Phi, z^*) + \sum_{z\in\Omega\setminus\{z^*\}} q(\Phi, z)} \\
&= \frac{q(\Psi, z^*) + \sum_{z\in\Omega\setminus\{z^*\}}\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}\right)\prod_{j=1}^n \Psi_{i_j}^{(z_j)}}{q(\Phi, z^*) + \sum_{z\in\Omega\setminus\{z^*\}}\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}\right)\prod_{j=1}^n \Phi_{i_j}^{(z_j)}} \\
&\leq \frac{q(\Psi, z^*) + \sum_{z\in\Omega_1}\epsilon\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)} + \sum_{S\in\Omega_2}\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*) + \sum_{z\in\Omega_2}(1-\epsilon)^m\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}.
\end{aligned}$$

From Lemma A.9 we know that we can separately bound

$$\frac{q(\Psi, z^*) + \sum_{z\in\Omega_1}\epsilon\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*)} \quad\text{and}\quad \frac{\sum_{z\in\Omega_2}\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}{\sum_{z\in\Omega_2}(1-\epsilon)^n\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}.$$

The second quantity is simply bounded above by $(1-\epsilon)^{-m} \leq \exp\left(\frac{\epsilon m}{1-\epsilon}\right) \leq \exp(2\epsilon m)$.

By properties of the gamma function, $\prod_{i=1}^K \Gamma(\alpha_i + r_i) \leq \Gamma(\alpha_0 + m)$ for any $r_1, \ldots, r_K \geq 0$ satisfying $r_1 + \cdots + r_K = m$. Since $\|\Phi - \Psi\|_{max} \leq \epsilon$ and $\Phi_{i_j}^{(z_j^*)} \geq c$ for all j, we have

$$\begin{aligned}
&\frac{q(\Psi, z^*) + \sum_{z\in\Omega_1}\epsilon\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}{q(\Phi, z^*)} \\
&= \frac{\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z^*))}{\Gamma(\alpha_i)}\right)\prod_{i=1}^m \Psi_i^{(z_i^*)} + \sum_{z\in\Omega_1}\epsilon\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z))}{\Gamma(\alpha_i)}}{\left(\prod_{i=1}^K \frac{\Gamma(\alpha_i+n_i(z^*)|)}{\Gamma(\alpha_i)}\right)\prod_{i=1}^m \Phi_i^{(z_i^*)}}
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{\prod_{i=1}^m \Psi_i^{(z_i^*)}}{\prod_{i=1}^m \Phi_i^{(z_i^*)}} + \frac{\epsilon|\Omega_1|\Gamma(m+\alpha_0)}{\left(\prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))\right)\left(\prod_{i=1}^m \Phi_i^{(z_i^*)}\right)} \\
&\le (1+\epsilon/c)^m + \frac{\epsilon|\Omega_1|\Gamma(m+\alpha_0)}{c^m \prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))} \\
&\le e^{\epsilon m/c} + \frac{\epsilon K^m \Gamma(m+\alpha_0)}{c^n \prod_{j=1}^K \Gamma(\alpha_j + n_j(z^*))}.
\end{aligned}
$$

Where the last line follows by observing that $|\Omega_1| \le |\Omega| = K^m$.

Additionally, by the log-convexity of Γ on the positive reals, we know that for positive $x_1, \ldots, x_K$, $\Gamma(x_1)\cdots\Gamma(x_K) \ge (\Gamma(x_1/K + \cdots x_K/K))^K$. Thus

$$
\frac{p(x \,|\, \Psi)}{p(x \,|\, \Phi)} \;\le\; e^{\epsilon m/c} + \frac{\epsilon K^m \Gamma(m+\alpha_0)}{c^m (\Gamma(\alpha_0/K + m/K))^K}.
$$

Taking logs and making use of $\epsilon < c/m$, we have

$$
\begin{aligned}
\ln \frac{p(x \,|\, \Psi)}{p(x \,|\, \Phi)} &\le \ln\left(e^{\epsilon m/c} + \frac{\epsilon \Gamma(\alpha_0 + m) K^m}{c^m (\Gamma(\alpha_0/K + m/K))^K}\right) \\
&\le \ln\left(1 + \frac{2m\epsilon}{c} + \frac{\epsilon \Gamma(\alpha_0 + m) K^m}{c^m (\Gamma(\alpha_0/K + m/K))^K}\right) \\
&\le \epsilon\left(\frac{2m}{c} + \left(\frac{K}{c}\right)^m \frac{\Gamma(\alpha_0 + m)}{(\Gamma(\alpha_0/K + m/K))^K}\right)
\end{aligned}
$$

By Gauss' multiplicative theorem and the log-convexity of Γ, we know for any positive integer k and any $a > 0$,

$$
\frac{\Gamma(ka)}{\Gamma(a)^k} \;\le\; \max(1, a^k) k^{ak-1/2}.
$$

Applying this to the above gives us the lemma statement. □

A.2.2 Proof of Theorem 3.7

Lemma 3.4. *Pick any $\delta > 0$ and any $\theta \in \Theta$ within δ of the optimal MAP solution for Z, that is,*

$$
\log q_Z(\theta) \ge \sup_{\theta' \in \Theta} \log q_Z(\theta') - \delta.
$$

Then the log-likelihood of any $\theta' \in \Theta$ can exceed that of θ by at most

$$\log p(X|\theta') - \log p(X|\theta) \leq \frac{1}{k} \left(\delta + \log q_0(\theta) - \log q_0(\theta')\right).$$

Proof. Note that since θ is within δ of the supremum of $\ln q_Z$, we have

$$-\delta \leq \ln q_Z(\theta) - \ln q_Z(\theta') = \ln \frac{q_0(\theta) p(X \mid \theta)^k}{q_0(\theta') p(X \mid \theta')^k} = \ln \frac{q_0(\theta)}{q_0(\theta')} - k \ln \frac{p(X \mid \theta')}{p(X \mid \theta)}.$$

Rearranging the above gives us

$$\ln p(X \mid \theta') - p(X \mid \theta) = \ln \frac{p(X \mid \theta')}{p(X \mid \theta)} \leq \frac{1}{k} \left(\delta + \ln \frac{q_0(\theta)}{q_0(\theta')}\right).$$

□

A.2.3 Proof of Theorem 3.8

Define the TM-MLE(α, K, m) problem to be the TM-MLE(α) problem where the number of topics is K and the number of words per document is bounded from above by m. TM-MAP(α, β, K, m) and TM-APPROX-SAMPLING(α, β, K, m) are defined analogously.

For every $c > 0$, define

$$S_c = \left\{\Psi \in \Delta^{V \times K} \,:\, \Psi \text{ is } c\text{-smooth}\right\}.$$

If m is the length of the longest document, then we have by Lemma 3.6 that the max-norm is $(g(K, m, c, \alpha), S_c)$-admissible for

$$g(K, m, c, \alpha) = \frac{2m}{c} + \max\left(1, \left(\frac{\alpha_0 + m}{K}\right)^K\right) \frac{K^{\alpha_0 + 2m - 1/2}}{c^m}.$$

The next thing we need to establish to apply our results from Sections 3.3 and 3.4 is that the prior distribution is well-behaved on neighborhoods of the maximum likelihood estimate. The following lemma gives us a handle on the Dirichlet distribution.

Lemma A.13. *Suppose that ν is the measure and q is the density associated with the symmetric Dirichlet distribution over Δ^N with parameter α. Then for any $\epsilon > 0$ and any point $x \in \Delta^N$ s.t. $\min_i x_i \geq \epsilon$ we have*

$$\log q(x) \;\geq\; -\operatorname{poly}(N, \alpha, 1/\alpha, 1/\epsilon)$$

which implies for any $x \in \Delta^N$

$$\log \nu(B_{\ell_2}(x, \epsilon)) \;\geq\; -\operatorname{poly}(N, \alpha, 1/\alpha, 1/\epsilon).$$

Further, if $\alpha \geq 1$, we have

$$\log q(x) \;\leq\; \operatorname{poly}(N, \alpha).$$

Proof. Recall that

$$q(x) \;=\; \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N}\, x_1^{\alpha-1} \cdots x_N^{\alpha-1}.$$

We consider two cases.

Case 1: $\alpha < 1$**.** In this case q is a convex probability density with minimum at $(1/N, \ldots, 1/N)$. Thus,

$$q(x) \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \cdot \left(\frac{1}{N}\right)^{N(\alpha-1)} \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N} \cdot N^{-N}.$$

Notice by Γ's recurrence relation

$$\Gamma(\alpha) = \frac{\Gamma(1+\alpha)}{\alpha} \leq \frac{1}{\alpha}$$

for $\alpha \in (0,1)$. Moreover, $\Gamma(t) \geq 3/4$ for any real $t > 0$. Thus, we have

$$q(x) \geq \frac{3}{4}\left(\frac{\alpha}{N}\right)^N \geq 2^{-\operatorname{poly}(N,1/\alpha)}.$$

Case 2: $\alpha \geq 1$. When $x_i \geq \epsilon$ for $i = 1, \dots, n$, we have

$$q(x) \geq \frac{\Gamma(N\alpha)}{\Gamma(\alpha)^N}\epsilon^{N(\alpha-1)} \geq 2^{-\operatorname{poly}(N,\alpha,1/\epsilon)}.$$

Then the inequalities dealing with the density q in the lemme statement can be gleaned from the above two cases.

Now we turn to lower bounding $\nu(B_{\ell_2}(x,\epsilon) \cap \Delta^N)$. First, note that $\operatorname{vol}(B_{\ell_2}(x,\epsilon) \cap \Delta^N)$ is minimized for $x \in \Delta^N$ when x is a corner of Δ^n. Thus we can consider $x \in \Delta^n$ such that w.l.o.g. $x_1 = 1$ and $x_i = 0$ for $i = 2, \dots, N$. We claim $B_{\ell_2}(x,\epsilon) \cap \Delta^N$ contains a regular simplex with edge length $\epsilon/2N$ satisfying that $\min_i x_i \geq \epsilon/2N$ for all $x \in S$. To see this, let S be the simplex created by the convex hull of $x^{(1)}, \dots, x^{(N)} \in \Delta^N$ where

$$x_i^{(1)} = \begin{cases} 1 - \frac{(N-1)\epsilon}{2N} & \text{if } i = 1 \\ \frac{\epsilon}{2N} & \text{o/w} \end{cases}$$

and

$$x_i^{(k)} = \begin{cases} 1 - \frac{\epsilon}{2N}\left(N - 2 + \frac{1+\sqrt{2}}{\sqrt{2}}\right) & \text{if } i = 1 \\ \frac{(1+\sqrt{2})\epsilon}{2\sqrt{2}n} & \text{if } i = k \\ \frac{\epsilon}{2N} & \text{o/w} \end{cases}$$

for $k = 2, \dots, N$. Then one can see that

- $x^{(k)} \in \Delta^N$ for $k = 1, \dots, N$,
- $\|x^{(k)} - x^{(k')}\| = \frac{\epsilon}{2N}$ for all $k \neq k'$,

- $x^{(k)} \in B_{\ell_2}(x, \epsilon)$ for $k = 1, \ldots, N$, and
- $x_i^{(k)} \geq \frac{\epsilon}{2N}$ for $i, k = 1, \ldots, N$.

Then the simplex S lying in the convex hull of $x^{(1)}, \ldots, x^{(N)}$ is a regular simplex with edge length $\epsilon/2N$ satisfying that $\min_i x_i \geq \epsilon/2N$ for all $x \in S$. Therefore for any $x \in \Delta^N$,

$$\nu(B_{\ell_2}(x, \epsilon) \cap \Delta^N) \geq \mathrm{vol}(S) \cdot \inf_{x \in S} q(x) = \frac{\sqrt{N+1}}{N! 2^{N/2}} \cdot \left(\frac{\epsilon}{2N}\right)^N \cdot \inf_{x \in S} q(x) \geq 2^{-\operatorname{poly}(N, \alpha, 1/\alpha, 1/\epsilon)}.$$

□

We are now ready to apply Theorem 3.7.

Theorem A.14. *Let $\alpha > 0$, $c = 1/V$, $\beta \geq 1$, $K, m \in \mathbb{N}$, and let Π_c denote the promise that $\Psi_{ML} \in S_c$, then Π_c-TM-MLE$(\alpha, K, m) \leq_P$ TM-MAP(α, β, K, m) where the reduction is polynomial in the input size and $(1/c)^m$, K^m, and $\max\{\beta, 1/\beta\}$.*

Proof. Suppose that q is the density associated with the symmetric Dirichlet distribution over Δ^V with parameter β. The prior density q_0 we are interested in is the product distribution, i.e. for $\Psi \in \mathbb{R}^{V \times K}$,

$$q_0(\Psi) = q(\Psi^{(1)}) \cdots q(\Psi^{(K)}).$$

From Lemma A.13, we know that q is bounded above by $2^{\operatorname{poly}(V, \beta)}$. The density q_0 of the product distribution is thus bounded above by $2^{\operatorname{poly}(V, K, \beta)}$.

From Lemma 3.6, we know that max-norm is $(g(K, m, c, \alpha), S_c)$-admissible. Thus, in order to apply Theorem 3.7, we need to show the existence of a topic matrix $\widehat{\Psi}$ satisfying the following three conditions. For small enough $\epsilon > 0$,

(a) $\widehat{\Psi} \in S_c$,

(b) $\|\widehat{\Psi} - \Psi_{ML}\|_{max} \leq \epsilon$, and

(c) $q_0(\widehat{\Psi}) \geq 2^{-\operatorname{poly}(V,K,\beta,1/\epsilon)}$

To construct such a $\widehat{\Psi}$, let us first denote $\Psi = \Psi_{ML}$ and let $s = \min(\epsilon, 1/V^2)$. Consider a particular column j. If it is the case that $\Psi_i^{(j)} \geq s$ for all rows i, then we take $\widehat{\Psi}^{(j)} = \Psi^{(j)}$. Otherwise, because $\Psi^{(j)}$ is a distribution over V words and sums to one, this implies that there exists a row i^* such that $\Psi_{i^*}^{(j)} \geq \frac{1}{V} + \frac{s}{V}$. Then we take

$$\widehat{\Psi}_i^{(j)} = \begin{cases} \Psi_i^{(j)} - \frac{s}{V} & \text{if } i = i^* \\ \Psi_i^{(j)} + \frac{s}{V(V-1)} & \text{otherwise} \end{cases}$$

Then $\widehat{\Psi}$ is a valid topic matrix. It is easy to check that it satisfies (a) and (b). To see (c), notice that $\widehat{\Psi}_i^{(j)} \geq \frac{s}{V(V-1)}$ for all i, j. By Lemma A.13, this implies that every column j satisfies $q(\widehat{\Psi}^{(j)}) \geq 2^{-\operatorname{poly}(V,\beta,1/\epsilon)}$, which implies $q_0(\Psi) \geq 2^{-\operatorname{poly}(V,K,\beta,1/\epsilon)}$. □

The ML estimate in the construction in Theorem 3.3 lies in S_c for $c = 1/V$. The construction also satisfies that $K = 2$ and $m = 2$ (and that α is a constant), which means that the dominating factor $g(K, m, c, \alpha)$ is bounded above by $\operatorname{poly}(V)$. Theorem 3.8 follows as an immediate corollary.

Theorem 3.8. *For any fixed $\alpha > 0$ and $\beta \geq 1$,* TM-MAP(α, β) *is NP-hard.*

A.3 Proofs from Section 3.4

Lemma 3.10. *Take any $\epsilon, \delta > 0$ and $X \in \mathcal{X}^n$. If Z is the sequence created by duplicating X k times for*

$$k \;\geq\; \frac{2}{\epsilon}\left(\log\left(\frac{1}{\delta} - 1\right) + \log\left(\frac{1 - \nu_0(B_{d_{p,X}}(\theta_{ML}, \epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ML}, \epsilon/2))}\right)\right)$$

then $\nu_Z(B_{d_{p,X}}(\theta_{ML}, \epsilon)) \geq 1 - \delta$.

Proof. For any measurable set B, we may write

$$\nu_Z(B) \;=\; \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}[\theta \in B]p(X|\theta)^k]}{\mathbb{E}_{\theta\sim\nu_0}[p(X|\theta)^k]}.$$

Thus,

$$\begin{aligned}
\frac{\nu_Z(B_{d_{p,X}}(\theta_{ML},\epsilon))}{\nu_Z(\Theta \setminus B_{d_{p,X}}(\theta_{ML},\epsilon))} \;&=\; \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \in B_{d_{p,X}}(\theta_{ML},\epsilon)p(X|\theta)^k]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \notin B_{d_{p,X}}(\theta_{ML},\epsilon))p(X|\theta)^k]} \\
&\geq\; \frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \in B_{d_{p,X}}(\theta_{ML},\epsilon/2))(e^{-\epsilon/2}p(X|\theta_{ML}))^k]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \notin B_{d_{p,X}}(\theta_{ML},\epsilon))(e^{-\epsilon}p(X|\theta_{ML}))^k]} \\
&\geq\; e^{k\epsilon/2}\frac{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \in B_{d_{p,X}}(\theta_{ML},\epsilon/2))]}{\mathbb{E}_{\theta\sim\nu_0}[\mathbf{1}(\theta \notin B_{d_{p,X}}(\theta_{ML},\epsilon))} \\
&=\; e^{k\epsilon/2}\frac{\nu_0(B_{d_{p,X}}(\theta_{ML},\epsilon/2))}{\nu_0(B_{d_{p,X}}(\theta_{ML},\epsilon))}
\end{aligned}$$

Note that if the above is greater than $1/\delta - 1$, we have

$$\nu_Z(B_{d_{p,X}}(\theta_{ML},\epsilon)) \;=\; \frac{\nu_Z(B_{d_{p,X}}(\theta_{ML},\epsilon))}{\nu_Z(B_{d_{p,X}}(\theta_{ML},\epsilon)) + \nu_Z(\Theta \setminus B_{d_{p,X}}(\theta_{ML},\epsilon))} \;\geq\; 1-\delta.$$

However, this condition is satisfied when

$$k \;\geq\; \frac{2}{\epsilon}\log\left(\frac{1}{\delta}-1\right) + \log\left(\frac{\nu_0(\Theta \setminus B_{d_{p,X}}(\theta_{ML},\epsilon))}{\nu_0(B_{d_{p,X}}(\theta_{ML},\epsilon/2))}\right). \qquad \square$$

A.3.1 Proof of Theorem 3.12

The following lemma shows that if two topic matrices are close in log-likelihood distance and one of them is smooth, then the other must also be smooth.

Lemma A.15. *Pick any $\epsilon > 0$. Suppose Ψ and Φ are two topic matrices. Let $\alpha \in \mathbb{R}^K$ be the parameter to our Dirichlet prior with $\alpha_i \geq \alpha_{min}$ for all i and $\alpha_0 = \sum \alpha_i$. If Ψ is c-smooth and $d_p(\Psi, \Phi) < \ln 1/\epsilon$, then Φ must be $\frac{\epsilon c \alpha_{min}}{\alpha_0}$-smooth.*

Proof. Suppose that Φ is not $\frac{\epsilon c \alpha_{min}}{\alpha_0}$-smooth. Then there exists a row i such that $\Phi_i^{(j)} \leq$

$\frac{\epsilon c \alpha_{min}}{K\alpha_0}$ for all $j = 1, \ldots, K$. Take x to be the document with a single instance of the word i in it. Then if $\theta \sim \text{Dirichlet}(\alpha)$,

$$\ln p(x \,|\, \Phi) = \ln \mathbb{E}_\theta \left[\sum_{j=1}^{K} \Phi_i^{(j)} \theta_j \right] = \ln \sum_{j=1}^{K} \Phi_i^{(j)} \frac{\alpha_j}{\alpha_0} \leq \ln \sum_{j=1}^{K} \frac{\epsilon\, c\, \alpha_{min}}{\alpha_0} \cdot \frac{\alpha_j}{\alpha_0} = \ln \frac{\epsilon\, c\, \alpha_{min}}{\alpha_0}.$$

On the other hand, we have that there exists a j' such that $\Phi_i^{(j')} \geq c$. Thus,

$$\ln p(x \,|\, \Psi) \;=\; \ln \sum_{j=1}^{K} \Phi_i^{(j)} \frac{\alpha_j}{\alpha_0} \;\geq\; \ln \frac{c\, \alpha_{min}}{\alpha_0}.$$

Putting the above together, we have

$$d_p(\Psi, \Phi) \;=\; \sup_{x'} |\ln p(x' \,|\, \Psi) - \ln p(x' \,|\, \Phi)| \;\geq\; \ln p(x \,|\, \Psi) - \ln p(x \,|\, \Phi) \;\geq\; \ln \frac{1}{\epsilon}.$$

Thus we have a contradiction. □

Much of the proof of Theorem 3.12 is similar to the proof of Theorem 3.8. One key difference is that we care about lower bounding the probability mass of balls with respect to the Dirichlet(β) distribution. Because the ℓ_∞ and ℓ_2 norms are related by a factor which is polynomial in the dimension, Lemma A.13 also implies that

$$\log \nu(B_{\ell_\infty}(x, \epsilon)) \;\geq\; -\operatorname{poly}(N, \beta, 1/\beta, 1/\epsilon)$$

for any $x \in \Delta^N$.

Theorem 3.12. *For any fixed $\alpha, \beta > 0$,* TM-APPROX-SAMPLING(α, β) *is NP-hard.*

Proof. We will reduce from an instance of TM-MLE(α) from Theorem 3.3. In order to apply Theorem 3.11, let S be the set of all $1/V$-smooth matrices, S' be the set of all $1/(2VK)$-smooth matrices, S'' be the set of all $1/(4VK)$-smooth matrices, and let d be the max-norm distance. Then

(i) if $\Psi \in S$ then $B_{d_p}(\Psi, \ln(2)) \subset S'$ (Lemma A.15),

(ii) if $\Psi \in S'$ then $B_d(\Psi, 1/(4VK)) \subset S''$ (max-norm distance),

(iii) d is $(\mathrm{poly}(V), S'')$-admissible (Lemma 3.6, $K = m = 2$, and α is a constant), and

(iv) for all $\epsilon > 0$ and all $\Psi \in S$, $\nu_0(B_d(\Psi, \epsilon)) \geq 2^{-\mathrm{poly}(V,K,\alpha,1/\alpha,1/\epsilon)}$ (Lemma A.13).

Then since instances from Theorem 3.3 satisfy that $\Psi_{ML} \in S$, Theorem 3.11 implies that TM-APPROX-SAMPLING(α, β) is NP-hard. □

A.4 Proofs from Section 3.5

A.4.1 Proof of Theorem 3.13

Theorem 3.13. *Let* Π *be the promise that there exists a low-order polynomial* $\rho(\cdot,\cdot,\cdot)$ *such that if* $\theta_{ML} = (\boldsymbol{\mu}^*, \boldsymbol{\pi}^*, \sigma^*)$ *is an optimal maximum likelihood solution and* $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ *satisfies* $d_p(\theta_{ML}, \theta) < 1$*, then*

(i) $\|\mu_j\| \leq \rho(n, d, k)$ *for all* j*,*

(ii) $\sigma^2 \geq 1/\rho(n, d, k)$*,*

(iii) $\pi_j > 0$ *for all* j*, and*

(iv) $\pi_j^* \geq 1/\rho(n, d, k)$ *for all* j*.*

Then Π-MLE-MOGS-SV(k) *is NP-hard for* $k \geq 2$*.*

Proof. We will use the same reduction as in Chapter 2, reducing from the k-means problem [2] in which there exist low-order polynomials $\alpha(\cdot)$ and $\beta(\cdot)$ such that

- For an instance containing n points, each point is unique and has dimension at most $\alpha(n)$, with individual coordinates taking values in $\{-1, 0, 1\}$.

- Any set of means with k-means cost within a factor of $1 + 1/\beta(n)$ induces an optimal k-means partition of the data.

Again, we pad the points with zeros until the dimension d satisfies

$$d \;\geq\; \max\{16\beta(n)\ln k, 2n\alpha(n)\sqrt{1 + 2\ln k}\}$$

and solve the resulting MOGS-SV with $b = 1$. By the proof of Theorem 2.1, this solves the original k-means instance. Now we need to demonstrate that conditions (i), (ii), (iii), and (iv) hold for any $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ satisfying $d_p(\theta, \theta_{ML}) \leq 1$.

Proof of (i) We know from the proof of Theorem 2.1 that if θ has log-likelihood on this data set within n of θ_{ML}, then the partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ induced by $\boldsymbol{\mu}$ is an optimal k-means partition of the data set. By a bias-variance decomposition, this implies

$$\begin{aligned}\Phi(\boldsymbol{\mu}) &= \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mathrm{mean}(\mathcal{X}_j)\|^2 + \sum_{j=1}^{k} |X_j| \|\mu_j - \mathrm{mean}(\mathcal{X}_j)\|^2 \\ &= \Phi_{OPT} + \sum_{j=1}^{k} |X_j| \|\mu_j - \mathrm{mean}(\mathcal{X}_j)\|^2\end{aligned}$$

where Φ_{OPT} is the optimal k-means cost of this data set. If $\|\mu_j\| \geq 2d$ for some j, then we have

$$\Phi(\boldsymbol{\mu}) \;\geq\; \Phi_{OPT} + d.$$

But Lemma 2.2 implies

$$nd_p(\theta_{ML}, \theta) \;\geq\; LL(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*, \sigma^*) - LL(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)) \;\geq\; \frac{nd}{2} \ln\left(\frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}}\right) - 2n \ln k.$$

Rearranging, we see

$$\Phi_{OPT} + d \;\leq\; \Phi(\boldsymbol{\mu}) \;\leq\; \Phi_{OPT}\left(1 + \frac{2}{d}(1 + 2\ln k)\right).$$

We know $\Phi_{OPT} \leq n\alpha(n)^2$, since this is the cost of taking the origin to be the only center. Thus, we have

$$d^2 \;\leq\; n\alpha(n)^2(1 + 2\ln k)$$

which is not possible by our choice of d. Therefore, we have $\|\mu_j\| < 2d$ for all j.

Proof of (ii) Taking $\sigma^2 = \gamma\frac{\Phi_{OPT}}{nd}$, Lemma 2.5 implies

$$nd_p(\theta_{ML}, \theta) \;\geq\; \frac{nd}{2}\ln\gamma + \frac{nd}{2\gamma} - \frac{nd}{2} - 2n\ln k \;\geq\; \frac{nd}{2}\left(\frac{1}{2\gamma} - 1\right) - 2n\ln k.$$

Rearranging, we see

$$\sigma^2 \geq \frac{\Phi_{OPT}}{2nd(1 + \frac{2}{d}(1 + 2\ln k))}.$$

Since all the data points are unique and at distance at least 1 from each other, there must be at least one mean in any optimal solution that lies at least at distance 1/2 from one of the data points. Thus $\Phi_{OPT} \geq 1/4$ and

$$\sigma^2 \geq \frac{1}{16nd(1 + \ln k)}.$$

Proof of (iii) Now let $(\mathcal{X}'_1, \ldots \mathcal{X}'_k)$ be the partition induced by $(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$, i.e.

$$X'_j = \{x \,:\, j = \operatorname*{argmax}_i \pi_i\, N(x; \mu_i, \sigma^2)\},$$

breaking ties arbitrarily. Let $\hat{\mu}_j = \text{mean}(\mathcal{X}'_j)$ for each non-empty cluster. Then we have

$$\begin{aligned} LL(\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma) &\leq n\ln k + \sum_{j=1}^{k}\sum_{x\in\mathcal{X}'_j} \ln N(x; \mu_j, \sigma^2) \\ &= n\ln k + \frac{nd}{2}\ln\frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^{k}\sum_{x\in\mathcal{X}'_j}\|x - \mu_j\| \end{aligned}$$

$$
\begin{aligned}
&\leq n \ln k + \frac{nd}{2} \ln \frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \|x - \hat{\mu}_j\| \\
&\leq n \ln k + \frac{nd}{2} \ln \frac{nd}{2\pi\Phi(\hat{\boldsymbol{\mu}})} - \frac{nd}{2}
\end{aligned}
$$

From this we see

$$nd_p(\theta, \theta_{ML}) \ \geq \ \frac{nd}{2} \ln \left(\frac{\Phi(\hat{\boldsymbol{\mu}})}{\Phi_{OPT}} \right) - 2n \ln k.$$

Rearranging, we have

$$\Phi(\hat{\boldsymbol{\mu}}) \leq \Phi_{OPT} \left(1 + \frac{1}{\beta(n)} \right).$$

This implies that $(\mathcal{X}'_1, \ldots \mathcal{X}'_k)$ is an optimal k-means partition, which implies that none of the $\mathcal{X}'_j$ are empty, meaning each mixing weight must be non-zero.

Proof of (iv) Now let $\theta_{ML} = (\boldsymbol{\mu}^*, \boldsymbol{\pi}^*, \sigma^*)$. By the proof of (iii), we know that the partition $(\mathcal{X}^*_1, \ldots \mathcal{X}^*_k)$ such that

$$X^*_j = \{x \, : \, j = \operatorname*{argmax}_i \pi^*_i \, N(x; \mu^*_i, \sigma^{*2})\}$$

must satisfy that X^*_j is non-empty for any j. Further, by the convergence of the EM algorithm, we know that for any j,

$$\pi^*_j \ = \ \frac{1}{n} \sum_{x \in \mathcal{X}} \frac{\pi^*_j N(x; \mu^*_j, \sigma^{*2})}{\sum_{i=1}^{k} \pi^*_i N(x; \mu^*_i, \sigma^{*2})} \ \geq \ \frac{1}{n} \sum_{x \in \mathcal{X}_j} \frac{\pi^*_j N(x; \mu^*_j, \sigma^{*2})}{\sum_{i=1}^{k} \pi^*_i N(x; \mu^*_i, \sigma^{*2})} \ \geq \ \frac{1}{kn}$$

□

A.4.2 Proof of Lemma 3.14

Recall that for two parameter vectors $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\sigma})$, their parameter distance is defined as

$$d(\theta, \hat{\theta}) \;=\; \max_i \left(\left| \ln \frac{\pi_i}{\hat{\pi}_i} \right|, |\sigma^2 - \hat{\sigma}^2|, \|\mu_i - \hat{\mu}_i\|^2 \right).$$

Lemma A.16. *Suppose* $\pi, \hat{\pi}, \sigma^2, \hat{\sigma}^2 > 0$ *and* $|\log \pi/\hat{\pi}|, |\sigma^2 - \hat{\sigma}^2|, \|\mu - \hat{\mu}\|^2 \leq \epsilon$ *and* $x \in \mathbb{R}^d$, *then*

$$\left| \log \frac{\pi\, N(x \,|\, \mu, \sigma^2)}{\hat{\pi}\, N(x \,|\, \hat{\mu}, \hat{\sigma}^2)} \right| \leq \epsilon \cdot \max \Bigg(1 + \frac{d}{2\sigma^2} + \frac{2\|x - \mu\| + \epsilon}{2\hat{\sigma}^2} + \frac{\|x - \mu\|^2}{2\sigma^2\hat{\sigma}^2} \\ 1 + \frac{d}{2\hat{\sigma}^2} + \frac{2\|x - \hat{\mu}\| + \epsilon}{2\sigma^2} + \frac{\|x - \hat{\mu}\|^2}{2\sigma^2\hat{\sigma}^2} \Bigg).$$

Proof. The proof consists of first demonstrating

$$\log \frac{N(x \,|\, \mu, \sigma^2)}{N(x \,|\, \hat{\mu}, \hat{\sigma}^2)} \leq \epsilon \left(\frac{d}{2\sigma^2} + \frac{2\|x - \mu\| + \epsilon}{2\hat{\sigma}^2} + \frac{\|x - \mu\|^2}{2\sigma^2\hat{\sigma}^2} \right)$$

and then demonstrating

$$\log \frac{N(x \,|\, \hat{\mu}, \hat{\sigma}^2)}{N(x \,|\, \mu, \sigma^2)} \leq \epsilon \cdot \left(\frac{d}{2\hat{\sigma}^2} + \frac{2\|x - \hat{\mu}\| + \epsilon}{2\sigma^2} + \frac{\|x - \hat{\mu}\|^2}{2\sigma^2\hat{\sigma}^2} \right).$$

Because the proofs are symmetric, we will only demonstrate the first inequality. To begin, note that we can write out the likelihood ratio as follows.

$$\begin{aligned} \frac{N(x \,|\, \mu, \sigma^2)}{N(x \,|\, \hat{\mu}, \hat{\sigma}^2)} &= \left(\frac{\hat{\sigma}^2}{\sigma^2} \right)^{d/2} \exp \left[\frac{\|x - \hat{\mu}\|^2}{2\hat{\sigma}^2} - \frac{\|x - \mu\|^2}{2\sigma^2} \right] \\ &\leq \left(1 + \frac{\epsilon}{\sigma^2} \right)^{d/2} \exp \left[\frac{(\|x - \mu\| + \|\hat{\mu} - \mu\|)^2}{2\hat{\sigma}^2} - \frac{\|x - \mu\|^2}{2\sigma^2} \right] \\ &\leq \exp \left[\frac{d\epsilon}{2\sigma^2} + \frac{\|x - \mu\|^2 + 2\epsilon\|x - \mu\| + \epsilon^2}{2\hat{\sigma}^2} - \frac{\|x - \mu\|^2}{2\sigma^2} \right] \end{aligned}$$

$$= \exp\left[\frac{d\epsilon}{2\sigma^2} + \frac{2\epsilon\|x-\mu\| + \epsilon^2}{2\hat{\sigma}^2} + \frac{\|x-\mu\|^2}{2}\left(\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2}\right)\right]$$
$$\le \exp\left[\frac{d\epsilon}{2\sigma^2} + \frac{2\epsilon\|x-\mu\| + \epsilon^2}{2\hat{\sigma}^2} + \frac{\epsilon\|x-\hat{\mu}\|^2}{2\sigma^2\hat{\sigma}^2}\right]$$

Taking logs and factoring out ϵ gives us the inequality. □

Given the above, the following lemma is immediate.

Lemma 3.14. *Let $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\sigma})$ be two parameter vectors satisfying $\pi_j, \hat{\pi}_j > 0$ for all j. Then $d_p(\theta, \hat{\theta}) \le d(\theta, \hat{\theta}) \operatorname{poly}(1/\sigma_i^2, 1/\hat{\sigma}_i^2, \|\mu_i\|^2, \|\hat{\mu}_i\|^2)$.*

A.4.3 Proof of Lemma 3.15

Before we prove Lemma 3.15, we need to bound quantities related to the Normal-Inverse-Gamma distribution and the Beta distribution.

Lemma A.17. *Fix $\alpha, \beta, n_0 > 0$ and $\mu_0 \in \mathbb{R}^d$. Let q and ν be the measure associated with the Normal-Inverse-Gamma distribution with these parameters. Then for any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$, we have*

$$-\operatorname{poly}(\alpha, \beta, n_0, d, \|\mu\|, \|\mu_0\|, \sigma^2) \ \le\ \log q(\mu, \sigma^2) \ \le\ \operatorname{poly}(\alpha, \beta, n_0, d).$$

Moreover, if $d((\mu, \sigma^2), (\hat{\mu}, \hat{\sigma}^2)) = \max\{\|\mu - \hat{\mu}\|, |\sigma^2 - \hat{\sigma}^2|\}$, then

$$\log B_d((\mu, \sigma^2), \epsilon) \ge -\operatorname{poly}(\alpha, \beta, n_0, d, \|\mu\|, \|\mu_0\|, \sigma^2, 1/\epsilon).$$

Proof. The density q can be written out as

$$q(\mu, \sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{x}\right)\left(\frac{n_0}{2\pi\sigma^2}\right)^{d/2} \exp\left(-\frac{\|\mu-\mu_0\|^2}{2\sigma^2/n_0}\right).$$

To see the upper bound on the density, note that the mode of this distribution occurs at $\mu = \mu_0$ and $\sigma^2 = \frac{\beta}{\alpha+d/2+1}$.

The lower bound on the density follows by noting that (i) $\|\mu - \mu_0\|^2 \leq 2\|\mu\|^2 + 2\|\mu_0\|^2$ and (ii) $\Gamma(x) \geq 3/4$ for all x and

$$\Gamma(x) \leq \begin{cases} x^x = 2^{x \log x} & \text{for } x > 1 \\ \frac{1}{x} & \text{for } x \leq 1 \end{cases}.$$

The lower bound on the measure follows by combining the lower bound on the density for any point in $B_d((\mu, \sigma^2), \epsilon)$ along with the volume of $B_d((\mu, \sigma^2), \epsilon)$. □

Lemma A.18. *Let $\gamma > 0$ and take ν be the measure and q be the density associated with the symmetric Beta(γ, γ) distribution. For $\theta = (w, 1-w)$, $\hat{\theta} = (\hat{w}, 1-\hat{w})$, let $d(\theta, \hat{\theta}) = \max\left(|\log(w/\hat{w})|, |\log((1-w)/(1-\hat{w}))|\right)$. If $w, 1-w \geq \delta > 0$, we have*

$$q(\theta) \geq 2^{-\operatorname{poly}(1/\gamma, \gamma, 1/\delta)}$$

and for $\epsilon \in (0, \gamma)$,

$$\nu(B_d(\theta, \epsilon)) \;\geq\; 2^{-\operatorname{poly}(1/\gamma, \gamma, 1/\epsilon, 1/\delta)}.$$

Proof. Writing out the density, we have

$$q(\theta) \;= \frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2} w^{\gamma-1}(1-w)^{\gamma-1}.$$

The bound on $q(\theta)$ follows from Lemma A.13. To see the lower bound on $\nu(B_d(\theta, \epsilon))$, assume w.l.o.g. that $w \leq 1/2$. For any $\hat{w} > 0$, if $|\log(w/\hat{w})| \leq \epsilon$, then $|\log((1-w)/(1-\hat{w}))| \leq 2\epsilon$. This implies

$$\mathcal{I} := \{\hat{\theta} = (\hat{w}, 1-\hat{w}) \,:\, e^{-\epsilon/2} w \leq \hat{w} \leq e^{\epsilon/2} w\} \;\subset\; B_d(\theta, \epsilon).$$

Then we have

$$\begin{aligned}
\nu(B_d(\theta,\epsilon)) &\geq \nu(\mathcal{I}) \\
&\geq (e^{\epsilon/2}w - e^{-\epsilon/2}w)\min_{\hat{\theta}\in\mathcal{I}} q(\hat{\theta}) \\
&\geq \frac{\delta\epsilon}{2}\min_{\hat{\theta}\in\mathcal{I}} q(\hat{\theta}) \\
&\geq 2^{-\operatorname{poly}(\gamma,1/\gamma,1/\delta,1/\epsilon)} \qquad \square
\end{aligned}$$

Given the above two lemmas, Lemma 3.15 follows immediately.

Lemma 3.15. *Let q and ν be the prior density and measure, respectively, for the Bayesian mixture of two spherical Gaussians generative model with parameters $\alpha, \beta, \gamma, \mu_0, n_0$. For any $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \sigma)$ and any $\epsilon > 0$, we have*

$$\log q(\theta) \;\geq\; -\operatorname{poly}(1/\pi_i, 1/\sigma, |\mu_i\|, d, n_0, \alpha, \beta, \gamma, \|\mu_0\|)$$

and

$$\log \nu(B_d(\theta,\epsilon) \;\geq\; -\operatorname{poly}(1/\pi_i, 1/\sigma, |\mu_i\|, d, 1/\epsilon, n_0, \alpha, \beta, \gamma, \|\mu_0\|).$$

Further, if $\gamma \geq 1$, we have

$$\log q(\theta) \;\leq\; \operatorname{poly}(d, n_0, \alpha, \beta, \gamma, \|\mu_0\|).$$

Appendix B

Supplementary material for Chapter 5

B.1 Proofs from Section 5.1

Lemma B.1. *Suppose $w = (w_1, \ldots, w_k)$, $\theta = (\theta_1, \ldots, \theta_k)$, $z = (z_1, \ldots, z_n)$, and $x = (x_1, \ldots, x_n)$ were generated as in* (5.1)*. Then the joint probability over labelings and points can be written as*

$$\Pr(x, z) = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha + n)\Gamma(\alpha)^k} \prod_{j=1}^{k} \Gamma(n_j(z) + \alpha) q(C_j(z)).$$

Proof.

$$\begin{aligned}
\Pr(x, z) &= \int_{\triangle_k} \int_{\Theta} \cdots \int_{\Theta} Dir_\alpha(\pi) \pi_1^{n_1(z)} \cdots \pi_k^{n_k(z)} \prod_{j=1}^{k} Q_\beta(\theta_j) P_{\theta_j}(C_j(z)) d\theta_1 \cdots d\theta_k d\pi \\
&= \left(\int_{\triangle_k} Dir_\alpha(\pi) \pi_1^{n_1(z)} \cdots \pi_k^{n_k(z)} d\pi \right) \prod_{j=1}^{k} \left(\int_{\Theta} Q_\beta(\theta_j) P_{\theta_j}(C_j(z)) d\theta_j \right) \\
&= \frac{\Gamma(k \cdot \alpha) \prod_{i=1}^{k} \Gamma(n_i(z) + \alpha)}{\Gamma(k \cdot \alpha + n) \prod_{i=1}^{k} \Gamma(\alpha)} \prod_{j=1}^{k} \left(\int_{\Theta} Q_\beta(\theta_j) P_{\theta_j}(C_j(z)) d\theta_j \right) \\
&= \frac{\Gamma(k \cdot \alpha)}{\Gamma(k \cdot \alpha + n)\Gamma(\alpha)^k} \prod_{j=1}^{k} \Gamma(n_j(z) + \alpha) q(C_j(z)).
\end{aligned}$$

□

Lemma 5.1. $\Pr(z_i = j \mid z_{-i}, x)$ *is proportional to* $(\alpha + n_j(z_{-i}))\Delta(C_j(z), i)$.

Proof. Using the definition of conditional probability and Lemma B.1, we have

$$\begin{aligned}
\Pr(z_i = j \mid z_{-i}, x) &\propto \Pr(z_i = j, z_{-i}, x) \\
&\propto \Gamma(n_j(z_{-i}) + \alpha + 1)q(C_j(z_{-i}) \cup \{i\}) \prod_{l \neq j} (\Gamma(n_l(z_{-i}) + \alpha)q(C_l(z_{-i}))) \\
&= \frac{\Gamma(n_j(z_{-i}) + \alpha + 1)q(C_j(z_{-i}) \cup \{i\})}{\Gamma(n_j(z_{-i}) + \alpha)q(C_j(z_{-i}))} \prod_{l=1}^{k} (\Gamma(n_l(z_{-i}) + \alpha)q(C_l(z_{-i}))) \\
&\propto (\alpha + n_j(z_{-i})) \cdot \frac{q(C_j(z) \cup \{i\})}{q(C_j(z) \setminus \{i\})} \\
&= (\alpha + n_j(z_{-i}))\Delta(C_j(z), i). \qquad \square
\end{aligned}$$

B.2 Proofs from Section 5.2

Lemma 5.4. *The state space* $\Omega_{\leq k}(n)$ *is isomorphic to the set of equivalence classes induced by* $\sim$ *over* $\{1, \ldots, k\}^n$, $\Omega^\sharp$. *Furthermore, the projected Gibbs sampler specified in Figure 5.2 is the exactly the chain induced by taking the equivalence classes of the states of the collapsed Gibbs sampler. Finally, projected Gibbs sampler is reversible with respect to*

$$\pi^\flat(\mathbb{C}) \propto \frac{1}{(k - |\mathbb{C}|)!} \prod_{S \in \mathbb{C}} \frac{\Gamma(|S| + \alpha)}{\Gamma(\alpha)} q(S).$$

Proof. Let P denote the collapsed Gibbs sampler and $P^\flat$ denote the projected Gibbs sampler.

Let us first show $\Omega_{\leq k}(n) \cong \Omega^\sharp$. To do this we will give a bijection $C : \Omega^\sharp \to \Omega_{\leq k}(n)$. Define C as

$$C([z]) = \{\{i : z_i = j\} : j \in \{1, \ldots, k\}\} \setminus \{\emptyset\}.$$

To see C is injective, let z, z' be labels such that $C([z]) = C([z'])$. We know there exists a $k' \leq k$ such that $C([z]) = \{S_1, \ldots, S_{k'}\}$. Let σ be a permutation such that for $i \in \{1, \ldots, n\}$, we have $\sigma(z)_i = j$ where $i \in S_j$ and let σ' be a permutation such that for $i \in \{1, \ldots, n\}$, we have $\sigma'(z')_i = j$ where $i \in S_j$. Since the set of permutations is a group under composition, we know that σ' has an inverse, call it $\bar{\sigma}$. Thus, by construction, we have $\bar{\sigma}(\sigma(z)) = z'$. Again, because the set of permutations is a group under composition, $\bar{\sigma} \circ \sigma$ is a permutation and $z \sim z'$.

To see that C is surjective, let $\mathbb{C} = \{S_1, \ldots, S_{k'}\}$ be a clustering in $\Omega_{\leq k}(n)$. Then consider the following labeling z: for each index i, if $i \in S_j$, then $z_i = j$. Since $\mathbb{C}$ is a clustering, every index is in some set $S \in \mathbb{C}$. Further, since $k' \leq k$, $z \in \{1, \ldots, k\}^n$. Finally, by construction, it is clear that $C(z) = \mathbb{C}$.

Now let z be a labeling and let $C = \mathbb{C}([z])$. How does P transition from z? With probability $1/n$, we choose an index i, and then we update according to $\Pr(z_i = j \mid z_{-i}, x)$. What is the probability that we move i to a cluster $S \in \mathbb{C}$? Say that cluster S has label j under z. Then,

$$\begin{aligned}
\Pr(\text{move } x \text{ to } S) &= \Pr(z_i = j \mid z_{-i}, x) \\
&\propto (\alpha + n_j(z_{-i})) \cdot \frac{q(C_j(z) \cup \{i\})}{q(C_j(z) \setminus \{i\})} \\
&= (\alpha + |S \setminus \{i\}|) \cdot \frac{q(S \cup \{i\})}{q(S \setminus \{i\})}.
\end{aligned}$$

Let E denote the event that i is moved to its own cluster, let $r = k - |\mathbb{C}|$, and let $a_1, \ldots, a_r$ be the empty labels of z_{-i}. Then,

$$\begin{aligned}
\Pr(E) &= \sum_{j=1}^{r} \Pr(z_i = a_j \mid z_{-i}, x) \\
&\propto \sum_{j=1}^{r} (\alpha + n_{a_j}(z_{-i})) \cdot \frac{q(C_{a_j}(z) \cup \{i\})}{q(C_{a_j}(z) \setminus \{i\})} \\
&= \sum_{j=1}^{r} \alpha \cdot q(\{i\}) \;\; = \;\; \alpha(k - |\mathbb{C}|)q(\{i\}).
\end{aligned}$$

Note that the above transition probabilities, $P^\flat(\mathbb{C}, \mathbb{C}') = P^\flat(z, C^{-1}(\mathbb{C}'))$, do not depend specifically on z but are equal for all $z' \in C^{-1}(\mathbb{C})$. Thus, by Lemma 5.2, $P^\flat$ is the induced Markov chain, $P^\sharp$, and $\pi^\flat = \pi^\sharp$. □

B.3 Proofs from Section 5.3

Lemma 5.6. *Let $\sigma^2, \mu_0, \sigma_0^2, Q_\beta, P_\theta, x$ be as given above. Then for any set of indices $S \subset \{1, \ldots, n\}$, we have $q(S) = L(S)R(S)$ where $L(S)$ is the probability assigned to S by the max-likelihood model,*

$$L(S) = \left(\frac{1}{2\pi\sigma^2}\right)^{|S|d/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i\in S}\|x_i - \mu(S)\|^2\right),$$

and $R(S)$ penalizes how far $\mu(S)$ is from μ_0:

$$R(S) = \left(\frac{\sigma^2}{\sigma^2 + |S|\sigma_0^2}\right)^{d/2} \exp\left(-\frac{|S|\|\mu_0 - \mu(S)\|^2}{2(\sigma^2 + |S|\sigma_0^2)}\right).$$

Proof. To facilitate our calculations, let $\tau = \frac{1}{\sigma^2}$ and $\tau_0 = \frac{1}{\sigma_0^2}$. Further

$$\begin{aligned}
q(S) &= \int_\Theta Q_\beta(\theta) P_\theta(S) d\theta \\
&= \int_{\mathbb{R}^d} \left(\frac{\tau}{2\pi}\right)^{nd/2} \exp\left[-\frac{\tau}{2}\left(\sum_{i\in S}\|x_i - \mu\|^2\right)\right] \left(\frac{\tau_0}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau_0}{2}\|\mu - \mu_0\|^2\right] d\mu \\
&= \left(\frac{\tau}{2\pi}\right)^{nd/2} \left(\frac{\tau_0}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau}{2}\left(\sum_{i\in S}\|x_i - \mu(S)\|^2\right)\right] \cdot \\
&\qquad \int_{\mathbb{R}^d} \exp\left[-\frac{1}{2}\left(\tau n\|\mu(S) - \mu\|^2 + \tau_0\|\mu - \mu_0\|^2\right)\right] d\mu \\
&= \left(\frac{\tau}{2\pi}\right)^{nd/2} \left(\frac{\tau_0}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau}{2}\left(\sum_{i\in S}\|x_i - \mu(S)\|^2\right)\right] \cdot \\
&\qquad \int_{\mathbb{R}^d} \exp\left[-\frac{1}{2}\left((n\tau + \tau_0)\left\|\mu - \frac{n\tau\mu(S) + \tau_0\mu_0}{n\tau + \tau_0}\right\|^2 + \frac{n\tau\tau_0}{n\tau + \tau_0}\|\mu(S) - \mu_0\|^2\right)\right] d\mu
\end{aligned}$$

$$= \left(\frac{\tau}{2\pi}\right)^{nd/2} \left(\frac{\tau_0}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau}{2}\left(\sum_{i \in S} \|x_i - \mu(S)\|^2 + \frac{n\tau_0}{n\tau + \tau_0}\|\mu(S) - \mu_0\|^2\right)\right] \cdot$$

$$\int_{\mathbb{R}^d} \exp\left[-\frac{1}{2}\left((n\tau + \tau_0)\left\|\mu - \frac{n\tau\mu(S) + \tau_0\mu_0}{n\tau + \tau_0}\right\|^2\right)\right] d\mu$$

$$= \left(\frac{\tau}{2\pi}\right)^{nd/2} \left(\frac{\tau_0}{n\tau + \tau_0}\right)^{d/2} \exp\left[-\frac{\tau}{2}\left(\sum_{i \in S} \|x_i - \mu(S)\|^2 + \frac{n\tau_0}{n\tau + \tau_0}\|\mu(S) - \mu_0\|^2\right)\right]$$

Substituting back in our identities for τ and τ_0, we can see that $q(S) = L(S)R(S)$. □

B.4 Proofs from Section 5.4

B.4.1 Proof of Theorem 5.8

Lemma B.2. *Say $x = (x_1, \dots, x_n)$ is a sequence of points in $\mathbb{R}^d$, $S \subset \{1, \dots, n\}$ a subset of indices, and $\epsilon > 0$.*

1. *If $i \in S$ and*

$$|S| \geq 1 + \frac{2\|x_i\|}{(\sqrt{1+\epsilon} - 1)\|x_i - \mu(S)\|}$$

we have $\|x_i - \mu_{S \setminus \{i\}}\|^2 \leq (1+\epsilon)\|x_i - \mu(S)\|^2$.

2. *If $i \in \{1, \dots, n\}$ and*

$$|S| \geq 1 + 2\sqrt{\frac{2}{\epsilon}}$$

we have $\|x_i - \mu_{S \setminus \{i\}}\|^2 \leq 2\|x_i - \mu(S)\|^2 + \epsilon\|x_i\|^2$.

3. *If $i \in \{1, \dots, n\}$ and*

$$|S| \geq \frac{\sigma^2\|\mu(S)\|}{(\sqrt{1+\epsilon} - 1)\|x_i - \mu(S)\|}$$

then $\|x_i - \mu_S\|^2 \geq (1-\epsilon)\|x_i - \mu(S)\|^2$.

Proof. We first prove (1). Recall

$$\mu(S \setminus \{i\}) = \frac{(|S| - 1)\mu(S) - x_i}{|S| - 1}.$$

Thus,

$$\begin{aligned}\|x_i - \mu_{S\setminus\{i\}}\| &= \left\| x_i - \mu(S \setminus \{i\}) \frac{\sigma_0^2(|S|-1)}{\sigma^2 + \sigma_0^2(|S|-1)} \right\| \\ &= \frac{\|x_i\sigma^2 + \sigma_0^2(|S|-1)(x_i - \mu(S \setminus \{i\})\|}{\sigma^2 + \sigma_0^2(|S|-1)} \\ &= \frac{\left\| x_i\sigma^2 + \sigma_0^2(|S|-1)\left(x_i - \mu(S) + \frac{x_i}{|S|-1}\right)\right\|}{\sigma^2 + \sigma_0^2(|S|-1)} \\ &\leq \frac{(\sigma^2 + \sigma_0^2)\|x_i\|}{\sigma^2 + \sigma_0^2(|S|-1)} + \frac{\sigma_0^2(|S|-1)\|x_i - \mu(S)\|}{\sigma^2 + \sigma_0^2(|S|-1)} \\ &\leq \|x_i - \mu(S)\| + \frac{2}{|S|-1}\|x_i\|.\end{aligned}$$

Applying the lower bound on $|S|$ and squaring both sides gives us (1). To prove (2), we first consider the case where $i \in S$ and see that

$$\begin{aligned}\|x_i - \mu_{S\setminus\{i\}}\|^2 &\leq 2\left(\frac{(\sigma^2 + \sigma_0^2)\|x_i\|}{\sigma^2 + \sigma_0^2(|S|-1)}\right)^2 + 2\left(\frac{\sigma_0^2(|S|-1)\|x_i - \mu(S)\|}{\sigma^2 + \sigma_0^2(|S|-1)}\right)^2 \\ &\leq 2\|x_i - \mu(S)\|^2 + 2\left(\frac{2}{|S|-1}\right)^2 \|x_i\|^2.\end{aligned}$$

Applying the lower bound on $|S|$ gives us the desired bound. Similarly, the above holds for the case where $i \notin S$ and

$$|S| \geq 2\sqrt{\frac{2}{\epsilon}} - 1.$$

This gives us (2). To prove (3), we calculate the following:

$$\begin{aligned}\|x_i - \mu_S\| &\geq \|x_i - \mu(S)\| - \|\mu(S) - \mu_S\| \\ &= \|x_i - \mu(S)\| - \|\mu(S)\| \left(1 - \frac{\sigma_0^2|S|}{\sigma^2 + \sigma_0^2|S|}\right) \\ &= \|x_i - \mu(S)\| - \sigma_S^2\|\mu(S)\| \\ &\geq \|x_i - \mu(S)\| - \frac{\sigma^2}{|S|}\|\mu(S)\|.\end{aligned}$$

Applying the lower bound on $|S|$ and squaring both sides gives us (2). □

Lemma B.3. *Let $S_1, \ldots, S_6$, δ, r, and A be as in Section 5.4.1. Further, let $k \in \{1, \ldots, 6\}$, $i \in S_k$ and $x = x_i$. Then $\|x - \mu(S_k)\|^2 \leq (\delta r)^2$ and for any $S' \neq S_k$,*

$$\|x - \mu(S')\|^2 \geq r^2(1-\delta)^2.$$

If $S_k \subset A$, then

$$\|x - \mu(A)\|^2 \leq r^2 \left(\sqrt{\frac{1}{2}} + \delta \right)^2.$$

If $S_k \not\subset A$, then

$$\|x - \mu(A)\|^2 \geq r^2(1-\delta)^2.$$

Lemma B.4. *Let $\delta \leq \frac{1}{32}$. Then there is a constant $n_0 = \Omega(\sigma^2)$ s.t. for $n \geq n_0$ and for $i \in A$, $j \in S_1$,*

$$\left.\begin{matrix} \dfrac{\|x_i - \mu_{A\setminus\{i\}}\|^2}{\sigma^2 + \sigma^2_{A\setminus\{i\}}} - \dfrac{\|x_i - \mu_{S_1}\|^2}{\sigma^2 + \sigma^2_{S_1}} \\ \dfrac{\|x_j - \mu_{S_1\setminus\{j\}}\|^2}{\sigma^2 + \sigma^2_{S_1\setminus\{j\}}} - \dfrac{\|x_j - \mu_{A}\|^2}{\sigma^2 + \sigma^2_{A}} \\ \dfrac{\|x_j - \mu_{S_1\setminus\{j\}}\|^2}{\sigma^2 + \sigma^2_{S_1\setminus\{j\}}} - \dfrac{\|x_j - \mu_{S_2}\|^2}{\sigma^2 + \sigma^2_{S_2}} \end{matrix}\right\} \leq -\frac{r^2}{4\sigma^2}$$

Proof. Consider the first inequality. Let $\epsilon_1 = 1/8$, $\epsilon_2 = 1/20$, and $\lambda = 1/100$. Then we know from Lemmas B.2 and B.3 that there exists a c_1 s.t. if $|A|, |S_1| \geq c_1$ and $|S_1| \geq 26 \geq \frac{1+\lambda}{\epsilon_2 - \lambda}$, then

$$\begin{aligned} \frac{\|x_i - \mu_{A\setminus\{i\}}\|^2}{\sigma^2 + \sigma^2_{A\setminus\{i\}}} - \frac{\|x_i - \mu_{S_1}\|^2}{\sigma^2 + \sigma^2_{S_1}} &\leq (1+\epsilon_1)\frac{\|x_i - \mu(A)\|^2}{\sigma^2 + \sigma^2_{A\setminus\{i\}}} - (1-\lambda)\frac{\|x_i - \mu(S_1)\|^2}{\sigma^2 + \sigma^2_{S_1}} \\ &\leq (1+\epsilon_1)\frac{\|x_i - \mu(A)\|^2}{\sigma^2} - (1-\epsilon_2)\frac{\|x_i - \mu(S_1)\|^2}{\sigma^2} \end{aligned}$$

$$\leq \frac{r^2}{\sigma^2}\left((1+\epsilon_1)\left(\sqrt{\frac{1}{2}}-\delta\right)^2-(1-\epsilon_2)(1-\delta)^2\right)$$
$$\leq -\frac{r^2}{4\sigma^2}.$$

Now consider the second inequality . If we take $\epsilon_1 = 1/5$, $\epsilon_2 = 1/2$, and $\lambda = 1/4$, then we know from Lemma B.2 and B.3 that there exists a c_2 s.t. if $|A|, |S_1| \geq c_2$ and $|A| \geq 5 \geq \frac{1+\lambda}{\epsilon_2-\lambda}$, then

$$\begin{aligned}\frac{\|x_j-\mu_{S_1\setminus\{j\}}\|^2}{\sigma^2+\sigma^2_{S_1\setminus\{j\}}}-\frac{\|x_j-\mu_A\|^2}{\sigma^2+\sigma^2_A} &\leq \frac{2\|x_j-\mu(S_1)\|^2+\epsilon_1\|x_j\|^2}{\sigma^2+\sigma^2_{S_1\setminus\{j\}}}-(1-\lambda)\frac{\|x_j-\mu(A)\|^2}{\sigma^2+\sigma^2_A}\\ &\leq \frac{2\|x_j-\mu(S_1)\|^2+\epsilon_1\|x_j\|^2}{\sigma^2}-(1-\epsilon_2)\frac{\|x_j-\mu(A)\|^2}{\sigma^2}\\ &\leq \frac{r^2}{\sigma^2}(2\delta^2+\epsilon_1-(1-\epsilon_2)(1-\delta)^2)\\ &\leq -\frac{r^2}{4\sigma^2}.\end{aligned}$$

Finally we consider the last inequality. Using the same ϵ_1. ϵ_2, and λ as in the second inequality, we know for the same c_2, if $|S_1|, |S_2| \geq c_2$ and $|S_2| \geq 5 \geq \frac{1+\lambda}{\epsilon_2-\lambda}$, then

$$\begin{aligned}\frac{\|x_j-\mu_{S_1\setminus\{j\}}\|^2}{\sigma^2+\sigma^2_{S_1\setminus\{j\}}}-\frac{\|x_j-\mu_{S_2}\|^2}{\sigma^2+\sigma^2_{S_2}} &\leq \frac{2\|x_j-\mu(S_1)\|^2+\epsilon_1\|x_j\|^2}{\sigma^2+\sigma^2_{S_1\setminus\{j\}}}-(1-\lambda)\frac{\|x_j-\mu(S_2)\|^2}{\sigma^2+\sigma^2_{S_2}}\\ &\leq 2\frac{\|x_j-\mu(S_1)\|^2+\epsilon_1\|x_j\|^2}{\sigma^2}-(1-\epsilon_2)\frac{\|x_j-\mu(S_2)\|^2}{\sigma^2}\\ &\leq \frac{r^2}{\sigma^2}(2\delta^2+\epsilon_1-(1-\epsilon_2)(1-\delta)^2)\\ &\leq -\frac{r^2}{4\sigma^2}.\end{aligned}$$

Note that Lemma B.2 gives an explicit form for the size of both c_1 and c_2. In the case of c_1, the ratios $\frac{\|x_i\|}{\|x_i-\mu(A)\|}$ and $\frac{\|\mu(S_1)\|}{\|x_i-\mu(S_1)\|}$ are constant, thus $c_1 \geq \Omega(\sigma^2)$ suffices. In the case of c_2, we have that $\frac{\|\mu(S_2)\|}{\|x_j-\mu(S_2)\|}$ is constant, and thus $c_2 \geq \Omega(\sigma^2)$ suffices. □

Theorem 5.8. *Let* $0 < \delta \leq 1/32$, $\alpha > 0$, $0 < \sigma \leq \sigma_0$, *and* $k = 3$. *Then there is a constant*

$n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ s.t. for $n \geq n_0$ the mixing rate of the projected Gibbs sampler with parameters α, σ, σ_0, and k over Ω is bounded below as $\tau_{mix} \geq \frac{1}{24} \cdot e^{\frac{r^2}{8\sigma^2}}$.

Proof. There are only four possible ways P can transition out of V: we can choose an index in A and move it to S_1 or S_2, we can choose an index in S_1 or S_2 and move it to A, we can choose an index in S_1 and move it to S_2, or we can choose an index in S_2 and move it to S_1. Say that the first event is E_1, the second event is E_2, and the union of the last two events is E_3.

Let us establish bounds on the probability first of these events, E_1. Since S_1 and S_2 play symmetric roles in this event, we can say for any index $i \in A$

$$\begin{aligned}\Pr(E_1) &\leq \frac{4}{3}\left(\frac{(\alpha+n)\Delta(S_1,i)}{(\alpha+4n-1)\Delta(A,i)}\right) \\ &= \frac{4}{3}\overbrace{\frac{(\alpha+n)}{(\alpha+4n-1)}}^{a}\left(\overbrace{\frac{\sigma^2+\sigma^2_{A\setminus\{i\}}}{\sigma^2+\sigma^2_{S_1}}}^{b}\right)^{d/2}\overbrace{\exp\left(\frac{1}{2}\left[\frac{\|x_i-\mu_{A\setminus\{i\}}\|^2}{\sigma^2+\sigma^2_{A\setminus\{i\}}}-\frac{\|x_i-\mu_{S_1}\|^2}{\sigma^2+\sigma^2_{S_1}}\right]\right)}^{c}.\end{aligned}$$

It is easy to see that we can find an $n_1 = \Theta(\alpha)$ s.t. $a \leq 1/2$. By Lemma 5.5 we know we can find an $n_2 = \Theta(d)$ s.t. $b \leq 2^{1/d}$. By Lemma B.4, we know that there is an $n_3 = \Theta(\sigma^2)$ s.t. $c \leq e^{-\frac{r^2}{8\sigma^2}}$. Thus, taking $n \geq n_0' \geq \max\{n_1, n_2, n_3\}$ we have

$$\Pr(E_1) \leq \frac{4}{3}e^{-\frac{r^2}{8\sigma^2}}$$

To bound E_2, note that since S_1 and S_2 play symmetric roles in this event, we can say for some index $i \in S_1$

$$\Pr(E_2) \leq \frac{2}{3}\left(\frac{(\alpha+4n)\Delta(A,i)}{(\alpha+n-1)\Delta(S_1,i)}\right)$$

$$= \frac{2}{3} \overbrace{\frac{(\alpha+4n)}{(\alpha+n-1)}}^{a} \left(\overbrace{\frac{\sigma^2 + \sigma^2_{S_1 \setminus \{i\}}}{\sigma^2 + \sigma^2_A}}^{b} \right)^{d/2} \overbrace{\exp\left(\frac{1}{2} \left[\frac{\|x_i - \mu_{S_1\setminus\{i\}}\|^2}{\sigma^2 + \sigma^2_{S_1\setminus\{i\}}} - \frac{\|x_i - \mu_A\|^2}{\sigma^2 + \sigma^2_A} \right] \right)}^{c}.$$

It is easy to see that we can find an $n'_1 = \Theta(\alpha)$ s.t. $a \leq 5$. By Lemma 5.5 we know we can find an $n'_2 = \Theta(d)$ s.t. $b \leq 2^{1/d}$. By Lemma B.4, we know that there is an $n'_3 = \Theta(\sigma^2)$ s.t. $c \leq e^{-\frac{r^2}{8\sigma^2}}$. Thus, taking $n \geq n''_0 \geq \max\{n'_1, n'_2, n'_3\}$ we have

$$\Pr(E_2) \;\leq\; \frac{10}{3} e^{-\frac{r^2}{8\sigma^2}}.$$

Now consider E_3. Since S_1 and S_2 play also symmetric roles in this event, we can say for some index $i \in S_1$

$$\Pr(E_3) \leq \frac{2}{3} \left(\frac{(\alpha+n)\Delta(S_2, i)}{(\alpha+n-1)\Delta(S_1, i)} \right)$$

$$= \frac{2}{3} \overbrace{\frac{(\alpha+n)}{(\alpha+n-1)}}^{a} \left(\overbrace{\frac{\sigma^2 + \sigma^2_{S_1 \setminus \{i\}}}{\sigma^2 + \sigma^2_{S_2}}}^{b} \right)^{d/2} \overbrace{\exp\left(\frac{1}{2} \left[\frac{\|x_i - \mu_{S_1\setminus\{i\}}\|^2}{\sigma^2 + \sigma^2_{S_1\setminus\{i\}}} - \frac{\|x_i - \mu_{S_2}\|^2}{\sigma^2 + \sigma^2_{S_2}} \right] \right)}^{c}.$$

It is easy to see that we can find an $n''_1 = \Theta(\alpha)$ s.t. $a \leq 2$. By Lemma 5.5 we know we can find an $n''_2 = \Theta(d)$ s.t. $b \leq 2^{1/d}$. By Lemma B.4, we know that there is an $n''_3 = \Theta(\sigma^2)$ s.t. $c \leq e^{-\frac{r^2}{8\sigma^2}}$. Thus, taking $n \geq n'''_0 \geq \max\{n''_1, n''_2, n''_3\}$ we have

$$\Pr(E_3) \;\leq\; \frac{4}{3} e^{-\frac{r^2}{8\sigma^2}}$$

Thus, if we take $n_0 = \max\{n'_0, n''_0, n'''_0\}$, the conductance of S is bounded as

$$\Phi(S) \;\leq\; \Pr(E_1) + \Pr(E_2) + \Pr(E_3) \;\leq\; 6e^{-\frac{r^2}{8\sigma^2}}.$$

Applying Theorem 4.2, the mixing rate of P is bounded below as

$$\tau_{mix} \;\geq\; \frac{1}{4\Phi^*} \;\geq\; \frac{1}{24}e^{\frac{r^2}{8\sigma^2}}. \qquad \square$$

B.4.2 Proof of Theorem 5.9

Lemma B.5. *Let $\{C_1, C_2\}$ be a 2-partition of S_3 and suppose $n \geq 2$ and $\alpha \geq 1$, then*

$$q(C_1)q(C_2) \geq \left(\frac{\sigma^2}{n\sigma_0^2}\right)^{d/2} \exp\left(-\frac{r^2}{\sigma_0^2}\right) q(S_3).$$

Proof. If $n \geq 2$, then

$$\left(\frac{n\sigma_0^2}{\sigma^2}\right)\left(\frac{n\sigma_0^2}{\sigma^2}+1\right)^{d/2} \geq \left(\frac{(n/2)\sigma_0^2}{\sigma^2}+1\right)^{d}.$$

From the bias-variance decomposition, we additionally have the following.

$$\sum_{x\in C_1}\|\mu(S_3)-x\|^2 = \sum_{x\in C_1}\|\mu(C_1)-x\|^2 + n_1\|\mu(S_3)-\mu(C_1)\|^2$$
$$\sum_{x\in C_2}\|\mu(S_3)-x\|^2 = \sum_{x\in C_2}\|\mu(C_2)-x\|^2 + n_2\|\mu(S_3)-\mu(C_2)\|^2$$

Combining the above with the identity $\sigma_0^2 \geq \sigma^2$, we have

$$q(C_1)q(C_2) \geq \left(\frac{\sigma^2}{n\sigma_0^2}\right)^{d/2} \exp(-r^2/\sigma_0^2)q(S_3). \qquad \square$$

Lemma 5.10. *For $n \geq 2$ and $\alpha \geq 1$,*

$$\frac{\pi(A)}{\pi(V)} \leq \frac{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{\alpha^2-\alpha}{n}+\frac{r^2}{\sigma_0^2}\right)\sigma_0^d}{\sigma^d n^{\alpha-d/2}}.$$

Proof. Let z denote the single element of A. Since $P(x,\cdot)$ is a probability measure for all

$x \in \Omega_{\leq 3}(X_G)$,

$$\frac{1}{\pi(V)} \sum_{x \in A, y \in V^c} \pi(x) P(x,y) \ = \ \frac{\pi(A)}{\pi(V)} \sum_{y \in V^c} P(z,y) \ \leq \ \frac{\pi(A)}{\pi(V)}.$$

If we let $\Omega_2(S_3)$ denote the set of non-empty partitions of S_3, then by the definition of π and from Lemma B.5,

$$\begin{aligned} \frac{\pi(A)}{\pi(V)} &\leq \frac{\frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} q(S_3)}{\sum_{\{C_1,C_2\} \in \Omega_2(S_3)} \frac{\Gamma(|C_1|+\alpha)}{\Gamma(\alpha)} q(C_1) \cdot \frac{\Gamma(|C_2|+\alpha)}{\Gamma(\alpha)} q(C_2)} \\ &\leq \frac{2\Gamma(\alpha)\Gamma(n+\alpha)(n\sigma_0^2)^{d/2} \exp\left(\frac{r^2}{\sigma_0^2}\right)}{\sigma^d \sum_{k=n(1/2-\sqrt{2}/4)}^{n(1/2+\sqrt{2}/4)} \binom{n}{k} \Gamma(k+\alpha)\Gamma(n-k+\alpha)} \\ &\leq \frac{2\Gamma(\alpha)(n+\alpha)^{\alpha-1}(n\sigma_0^2)^{d/2} \exp\left(\frac{r^2}{\sigma_0^2}\right)}{\sigma^d \sum_{k=n(1/2-\sqrt{2}/4)}^{n(1/2+\sqrt{2}/4)} (k+1)^{\alpha-1}(n-k+1)^{\alpha-1}}. \end{aligned}$$

For $k \in [n(1/2 - \sqrt{2}/4), n(1/2 + \sqrt{2}/4)]$ and $n \geq 2$,

$$\begin{aligned} (k+1)^{\alpha-1}(n-k+1)^{\alpha-1} &\geq k^{\alpha-1}(n-k)^{\alpha-1} \\ &\geq (n(1/2-\sqrt{2}/4))^{\alpha-1}(n(1/2+\sqrt{2}/4))^{\alpha-1} \\ &\geq \left(\frac{1}{8}\right)^{\alpha-1} n^{2(\alpha-1)}. \end{aligned}$$

Additionally, for all $n, \alpha > 0$,

$$\frac{(n+\alpha)^{\alpha-1}}{n^{\alpha-1}} = \left(1 + \frac{\alpha}{n}\right)^{\alpha-1} \leq \exp\left(\frac{\alpha^2 - \alpha}{n}\right).$$

Thus,

$$\begin{aligned}\frac{\pi(A)}{\pi(V)} &\leq \frac{2^{3\alpha-3/2}\Gamma(\alpha)(n+\alpha)^{\alpha-1}}{\sqrt{2}\cdot n^{2\alpha-1}} \cdot \left(\frac{n\sigma_0^2}{\sigma^2}\right)^{d/2} \exp\left(\frac{r^2}{\sigma_0^2}\right) \\ &\leq \frac{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{\alpha^2-\alpha}{n}+\frac{r^2}{\sigma_0^2}\right)}{n^{\alpha-d/2}} \cdot \left(\frac{\sigma_0}{\sigma}\right)^d. \end{aligned}$$ □

Lemma B.6. *Let S_1, S_2, S_3 be as in Section 5.4.2. Further, let $i \in S_1 \cup S_2$ and $j \in S_3$. Then we have the following*

1. $\|x_i - \mu(S_1 \cup S_2)\|^2 \leq r^2\left(\frac{1}{2}+\delta\right)^2$,
2. $\|x_i - y\|^2 \geq r^2(1-2\delta)^2$, *and*
3. $\|x_i - \mu(S_1 \cup S_2)\|^2 \geq r^2\left(\sqrt{\frac{3}{4}}-\delta\right)^2$.
4. *For* $b \in [1/2, 1]$, $\|x_i - bx_j\|^2 \geq r^2\left(\sqrt{\frac{7}{12}}-2\delta\right)^2$. *(Law of Cosines)*

The above lemma follows from simple geometric arguments.

Lemma B.7. *Let $C \subset S_3$. For $\delta \leq \frac{1}{4}\left(\sqrt{\frac{7}{3}}-\frac{3}{2}\right)$, there exists an $n_0 = \Omega(\sigma^2)$ s.t. for $n \geq n_0$ if $i \in S_3$ and $|C| \geq n/2$,*

$$\frac{\|x_i - \mu_{C\setminus\{i\}}\|^2}{\sigma^2+\sigma^2_{C\setminus\{i\}}} - \frac{\|x_i - \mu_{S_1\cup S_2}\|^2}{\sigma^2+\sigma^2_{S_1\cup S_2}} \leq -\frac{r^2}{48\sigma^2}$$

and if $i \in S_1 \cup S_2$ and $|C| \geq 1$,

$$\frac{\|x_i - \mu_{S_1\cup S_2\setminus\{i\}}\|^2}{\sigma^2+\sigma^2_{S_1\cup S_2\setminus\{i\}}} - \frac{\|x_i - \mu_C\|^2}{\sigma^2+\sigma^2_C} \leq -\frac{r^2}{48\sigma^2}.$$

Proof. Consider the first inequality. Let L_1 denote the left-hand side and $\epsilon_1 = 1/3$, $\epsilon_2 = 1/2$, and $\lambda = 1/4$. We know from Lemmas B.2 and B.6 that there exists a c_1 s.t. if

$|C|, |S_1 \cup S_2| \geq c_2$ and if $|S_1 \cup S_2| \geq 5 \geq \frac{1+\lambda}{\epsilon_2 - \lambda}$, then

$$\begin{aligned} L_1 &\leq \frac{2\|x_i - \mu(C)\|^2 + \epsilon_1 \|x_i\|^2}{\sigma^2 + \sigma^2_{C\setminus\{i\}}} - (1-\lambda)\frac{\|x_i - \mu(S_1 \cup S_2)\|^2}{\sigma^2 + \sigma^2_{S_1 \cup S_2}} \\ &\leq \frac{2\|x_i - \mu(C)\|^2 + \epsilon_1 \|x_i\|^2}{\sigma^2 + \sigma^2_{C\setminus\{i\}}} - (1-\epsilon_2)\frac{\|x_i - \mu(S_1 \cup S_2)\|^2}{\sigma^2} \\ &\leq \frac{r^2}{\sigma^2}\left(2\delta^2 + \epsilon_1 - (1-\epsilon_2)\left(\sqrt{\frac{3}{4}} - \delta\right)^2\right) \leq -\frac{r^2}{48\sigma^2}. \end{aligned}$$

Now we turn our attention to the second inequality and let L_2 denote the left-hand side. We first observe that since $\sigma_0^2 \geq \sigma^2$ and $|C| \geq 1$, we have

$$\frac{\|x_i - \mu_C\|^2}{\sigma^2 + \sigma_C^2} \geq \frac{\left\|x_i - \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}\mu(C)\right\|^2}{2\sigma^2} \geq \frac{\left\|x_i - \frac{1}{2}\mu(C)\right\|^2}{2\sigma^2}$$

Take $\epsilon = 1/100$, then by Lemmas B.2 and B.6 that there exists a c_2 s.t. if $|S_1 \cup S_2| \geq c_2$, then

$$\begin{aligned} L_2 &\leq \frac{(1+\epsilon)\|x_i - \mu(S_1 \cup S_2)\|^2}{\sigma^2 + \sigma^2_{S_1 \cup S_2 \setminus\{i\}}} - \frac{\left\|x_i - \frac{1}{2}\mu(C)\right\|^2}{2\sigma^2} \\ &\leq \frac{r^2}{\sigma^2}\left((1+\epsilon)\left(\frac{1}{2} + \delta\right)^2 - \frac{1}{2}\left(\sqrt{\frac{7}{12}} - 2\delta\right)^2\right) \leq -\frac{r^2}{48\sigma^2}. \end{aligned}$$

Note that Lemma B.2 gives an explicit form for the size of both c_1 and c_2. In the case of c_1, the ratio $\frac{\|\mu(S_1 \cup S_2)\|}{\|x_i - \mu(S_1 \cup S_2)\|}$ is constant, thus $c_1 \geq \Omega(\sigma^2)$ suffices. In the case of c_2, we have that $\frac{\|x_i\|}{\|x_i - \mu(S_1 \cup S_2)\|}$ is constant, and thus $c_2 \geq \Omega(\sigma^2)$ suffices. □

Lemma 5.11. *For $\delta \leq \frac{1}{4}\left(\sqrt{\frac{7}{3}} - \frac{3}{2}\right)$, there exists an $n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ s.t. for $n \geq n_0$,*

$$\frac{1}{\pi(V)} \sum_{x \in B, y \in V^c} \pi(x) P(x, y) \leq 6 \exp\left(-\frac{r^2}{96\sigma^2}\right).$$

Proof. Let $\mathbb{C} = \{C_1, C_2, S_1 \cup S_2\}$ be a 3-partition in B, where C_1 and C_2 are non-empty.

To move from $\mathbb{C}$ to V^c, there are 2 possibilities.

1. Move $i \in S_3$ to $S_1 \cup S_2$ (Event E_1).
2. Move $i \in S_1 \cup S_2$ to C_1 or C_2 (Event E_2).

We want to bound from above the probability of each of these events occuring. Let us first handle the first event. We know that one of C_1 and C_2 must have cardinality greater than or equal to $n/2$. Since they each play symmetric roles, assume without loss of generality that $|C_1| \geq n/2$. Then we can establish the following:

$$\begin{aligned}\Pr(E_1) &\leq \Pr(\text{choose } i \text{ in } S_3)\frac{\alpha+2n}{\alpha+|C_1|-1}\cdot\frac{\Delta(S_1\cup S_2,i)}{\Delta(C_1,i)}\\ &\leq \frac{1}{3}\overbrace{\frac{\alpha+2n}{\alpha+n/2-1}}^{a}\left(\overbrace{\frac{\sigma^2+\sigma^2_{C_1\backslash\{i\}}}{\sigma^2+\sigma^2_{S_1\cup S_2}}}^{b}\right)^{d/2}\exp\left(\frac{1}{2}\left[\overbrace{\frac{\|x_i-\mu_{C_1\backslash\{i\}}\|^2}{\sigma^2+\sigma^2_{C_1\backslash\{i\}}}-\frac{\|x_i-\mu_{S_1\cup S_2}\|^2}{\sigma^2+\sigma^2_{S_1\cup S_2}}}^{c}\right]\right)\end{aligned}$$

We know by Lemmas 5.5 and B.7 that there exists an n_1 s.t. for $n \geq n_1$, we have $a \leq 5$, $b \leq 2^{1/d}$, and $c \leq -\frac{r^2}{48\sigma^2}$. Thus, $\Pr(E_1) \leq \frac{10}{3}e^{-\frac{r^2}{96\sigma^2}}$.

Now consider event E_2. Without loss of generality, say that moving i to C_1 has higher probability than moving i to C_2. We can do the same computations as above to establish that there exists an n_2 s.t. for $n \geq n_2$ such that $\Pr(E_2)$ is bounded above by

$$\begin{aligned}&\frac{4}{3}\cdot\frac{\alpha+n}{\alpha+2n-1}\left(\frac{\sigma^2+\sigma^2_{S_1\cup S_2\backslash\{i\}}}{\sigma^2+\sigma^2_{C_1}}\right)^{d/2}\exp\left(\frac{1}{2}\left[\frac{\|x_i-\mu_{S_1\cup S_2\backslash\{i\}}\|^2}{\sigma^2+\sigma^2_{S_1\cup S_2\backslash\{i\}}}-\frac{\|x_i-\mu_{C_1}\|^2}{\sigma^2+\sigma^2_{C_1}}\right]\right)\\ &\leq \frac{8}{3}e^{-\frac{r^2}{96\sigma^2}}\end{aligned}$$

Combining the above, we can conclude

$$\sum_{\mathbb{C}\in B,\mathbb{C}'\in V^c}\frac{\pi(\mathbb{C})P(\mathbb{C},\mathbb{C}')}{\pi(V)} \;\leq\; \left(\frac{10}{3}+\frac{8}{3}\right)\exp\left(-\frac{r^2}{96\sigma^2}\right) \;\leq\; 6\exp\left(-\frac{r^2}{96\sigma^2}\right).$$

Finally, both n_1 and n_2 satisfy $\Omega(\max\{\alpha, \sigma^2, d\})$. Taking $n_0 = \max\{n_1, n_2\}$ gives us the lemma statement. $\square$

Appendix C

Supplementary material for Chapter 6

C.1 Proofs from Section 6.3

Theorem 6.1. *Let $c_1, c_2 \geq 0$ such that $c_1 c_2 < 1$, $P^{(v)}$ is c_1-contractive, and $P^{(h)}$ is c_2-contractive. Then the mixing rate of the Gibbs sampler is bounded as*

$$\tau(\epsilon) \ \leq \ 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{C}{\epsilon}\right)$$

where $C = \min\left(\frac{\gamma_v^{(max)}}{\gamma_v^{(min)}}, \frac{\gamma_h^{(max)}}{\gamma_h^{(min)}}, \frac{c_2\gamma_v^{(max)}}{\gamma_h^{(min)}}\right)$.

Proof. To see $\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1c_2)} \log\left(\frac{1}{\epsilon}\min\left(\frac{\gamma_h^{(\max)}}{\gamma_h^{(\min)}}, \frac{c_2\gamma_v^{(\max)}}{\gamma_h^{(\min)}}\right)\right)$, we will use the same coupling (X_t, Y_t) as given in the first part of the proof. Then by similar arguments,

$$\begin{aligned}
\Pr(X_t \neq Y_t) &\leq \Pr(d_h(X_t, Y_t) \geq \gamma_h^{(\min)}) \\
&\leq \frac{\mathbb{E}[d_h(X_t, Y_t)]}{\gamma_h^{(\min)}} \\
&\leq \frac{(c_1c_2)^{t-1}\mathbb{E}[d_h(X_1, Y_1)]}{\gamma_h^{(\min)}} \\
&\leq \frac{(c_1c_2)^{t-1}\min(\gamma_h^{(\max)}, c_2\gamma_v^{(\max)})}{\gamma_h^{(\min)}}
\end{aligned}$$

Taking $t \geq 1 + \frac{1}{\log(c_1 c_2)} \log\left(\frac{\min(\gamma_h^{(\max)}, c_2\gamma_v^{(\max)})}{\gamma_v^{(\min)}}\right)$ makes the above less than ϵ. Applying Lemma 4.4 completes the proof. □

Lemma 6.2. *$P_{RBM}^{(v)}$ and $P_{RBM}^{(h)}$ are $\frac{\|W\|_1}{2}$- and $\frac{\|W^T\|_1}{2}$-contractive, respectively.*

Proof. Let $x, y \in \Omega$ be two configurations. We will prove the claim for the visible conditional distributions. The proof for the hidden conditional distributions will follow symmetrically.

For each visible node v_i, let $(X(v_i), Y(v_i))$ be the maximal coupling of $P^{(v)}(X(v_i) \,|\, x(h))$ and $P^{(v)}(Y(v_i) \,|\, y(h))$ guaranteed in Lemma 4.3. By doing this independently for all visible nodes, we have a valid coupling (X, Y) of $P^{(v)}(\cdot \,|\, x(h))$ and $P^{(v)}(\cdot \,|\, y(h))$. Then we can work out the expected Hamming distance of X and Y as

$$\begin{aligned}
\mathbb{E}[d_v(X,Y)] &= \sum_{i=1}^{n} \left\| P^{(v)}(X(v_i) \,|\, x(h)) - P^{(v)}(Y(v_i) \,|\, y(h)) \right\|_{TV} \\
&= \sum_{i=1}^{n} \left| P^{(v)}(X(v_i) = 1 \,|\, x(h)) - P^{(v)}(Y(v_i) = 1 \,|\, y(h)) \right| \\
&= \sum_{i=1}^{n} \left| \frac{1}{1 + \exp\left(-a_i - \sum_{j=1}^{m} W_{ij} x(h_j)\right)} - \frac{1}{1 + \exp\left(-a_i - \sum_{j=1}^{m} W_{ij} y(h_j)\right)} \right| \\
&\leq \sum_{i=1}^{n} \left| \frac{1 - \exp\left(\sum_{j=1}^{m} W_{ij}\left(y(h_j) - x(h_j)\right)\right)}{1 + \exp\left(\sum_{j=1}^{m} W_{ij}\left(y(h_j) - x(h_j)\right)\right)} \right| \\
&= \sum_{i=1}^{n} \left| \tanh\left(\frac{\sum_{j=1}^{m} W_{ij}\left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right)}{2}\right) \right| \\
&\leq \sum_{i=1}^{n} \frac{1}{2} \left| \sum_{j=1}^{m} W_{ij}\left(y(h_j) - x(h_j)\right) \right| \\
&\leq \frac{1}{2} \sum_{j \,:\, y(h_j) \neq x(h_j)} \sum_{i=1}^{n} |W_{ij}| \\
&\leq \frac{1}{2} \|W\|_1 d_h(x,y).
\end{aligned}$$

□

Lemma 6.5. *$P_S^{(h)}$ and $P_S^{(v)}$ are $\frac{1}{2}\|W^T\|_1$- and $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, respectively.*

Proof. We will first show that $P_S^{(h)}$ is $\frac{1}{2}\|W^T\|_1$-contractive. To do so, let $x, y \in \Omega$ be two configurations. Our coupling (X, Y) of $P_S^{(h)}(\cdot \,|\, x(v))$ and $P_S^{(h)}(\cdot \,|\, y(v))$ is exactly the same as the coupling given in the proof of Lemma 6.2. Then, from the proof of Lemma 6.2, we have

$$\begin{aligned}
\mathbb{E}[d_h(X,Y) \,|\, x(v), y(v)] &= \sum_{j=1}^{m} \left| P_S^{(h)}(X(h_j) = 1 \,|\, x(v)) - P_S^{(h)}(Y(h_j) = 1 \,|\, y(v)) \right| \\
&\leq \left| \frac{1 - \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ij}^{(k)}(\mathbf{1}[y(v_i) = k] - \mathbf{1}[x(v_i) = k])\right)}{1 + \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ij}^{(k)}(\mathbf{1}[y(v_i) = k] - \mathbf{1}[x(v_i) = k])\right)} \right| \\
&= \left| \tanh\left(\frac{\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ij}^{(k)}(\mathbf{1}[y(v_i) = k] - \mathbf{1}[x(v_i) = k])}{2} \right) \right| \\
&\leq \frac{1}{2} \left| \sum_{i=1}^{n}\sum_{k=1}^{K} W_{ij}^{(k)}(\mathbf{1}[y(v_i) = k] - \mathbf{1}[x(v_i) = k]) \right| \\
&\leq \frac{1}{2} \sum_{i \,:\, x(v_i) \neq y(v_i)} \sum_{j=1}^{m} |W_{ij}| \\
&\leq \frac{1}{2}\|W^T\|_1 d_v(x, y).
\end{aligned}$$

To prove $P_S^{(v)}$ is $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, we will again use Lemma 4.3 to construct independent couplings $(X(v_i), Y(v_i))$ of $P_S^{(v)}(v_i \,|\, x(h))$ and $P_S^{(v)}(v_i \,|\, y(h))$ for each visible node v_i. Then by Lemma 4.3, we have

$$\begin{aligned}
\mathbb{E}[d_v(X,Y) \,|\, x(h), y(h)] &= \sum_{i=1}^{n} \|P_S^{(v)}(X(v_i) \,|\, x(h)) - P_S^{(v)}(Y(v_i) \,|\, y(h))\|_{TV} \\
&= \sum_{i=1}^{n} \frac{1}{2} \sum_{k=1}^{K} |P_S^{(v)}(X(v_i) = k \,|\, x(h)) - P_S^{(v)}(Y(v_i) = k \,|\, y(h))| \\
&\leq \frac{1}{2} \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{k' \neq k} \left| \tanh\left(\frac{\sum_{j=1}^{m}(W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j))}{2} \right) \right| \\
&\leq \frac{1}{2} \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{k' \neq k} \frac{1}{2} \left| \sum_{j=1}^{m}(W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j) \right|
\end{aligned}$$

$$\leq \frac{1}{2} \sum_{j\,:\,x(h_j)\neq y(h_j)} \frac{K(K-1)}{2} \sum_{i=1}^{n} |W_{ij}|$$
$$= \frac{1}{2}\binom{K}{2} \|W\|_1 d_h(x,y). \qquad \square$$

C.2 Proofs from Section 6.4

Lemma 6.8. *Let $c_1, c_2, \epsilon_0, \delta_0, M > 0$ such that $c_1c_2 < 1$, $P^{(h)}$ is c_1-contractive, $P^{(v)}$ is c_2-contractive and $(\epsilon_0, \delta_0, M)$-gamble admissible. There exists a Markovian coupling (X_t, Y_t) such that if $\mathbb{E}[d_v(X_0, Y_0)] \leq M$, then for any $\delta > 0$, if*

$$t \;\geq\; \frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)} \log\left(\frac{2c_1M}{\delta\epsilon_0} \cdot \frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $\Pr(X_t(v) \neq Y_t(v)) \leq \delta$.

Proof. Let (X_s, Y_s) be the *interleaved coupling* whose initial state is (X_0, Y_0) and is evolved according to the following rule.

1. Draw $(X_{s+1}(h), Y_{s+1}(h))$ according to the c_1-contractive coupling of $P^{(h)}(\cdot\,|\,X_s(v))$ and $P^{(h)}(\cdot\,|\,Y_s(v))$.

2. If $d_h(X_{s+1}, Y_{s+1}) \leq \epsilon_0$, draw $(X_{s+1}(v), Y_{s+1}(v))$ according to the $(\epsilon_0, \delta_0, M)$-gamble coupling of $P^{(v)}(\cdot\,|\,X_{s+1}(h))$ and $P^{(v)}(\cdot\,|\,Y_{s+1}(h))$. Otherwise, draw $(X_{s+1}(v), Y_{s+1}(v))$ according to the c_2-contractive coupling of $P^{(v)}(\cdot\,|\,X_{s+1}(h))$ and $P^{(v)}(\cdot\,|\,Y_{s+1}(h))$.

It is not too hard to see that (X_s, Y_s) is a Markovian coupling of the alternating Gibbs sampler.

Let us define two stochastic processes $Z_s = d_h(X_{s+1}, Y_{s+1})$, and $S_i = \inf\{s > S_{i-1} : Z_s \leq \epsilon_0\}$ where $S_0 = 0$. Due to the definition of the interleaved coupling, it is not hard to see that for any finite $i \geq 1$, $S_i < \infty$ with probability one. Moreover, because of the Markovian nature of S_i, we know that given S_{i-1}, S_i is independent of $S_0, S_1, \ldots, S_{i-2}$.

Now let $T, K \geq 1$ be given. Then we can work out the following

$$\Pr(X_{KT}(v) \neq Y_{KT}(v)) \leq \overbrace{\sum_{k=1}^{K} \Pr(S_k \geq kT \,|\, S_{k-1} \leq (k-1)T - 1)}^{a} \\ + \overbrace{\Pr(X_{KT} \neq Y_{KT} \,|\, S_1 \leq T-1, \ldots S_K \leq KT - 1)}^{b}$$

We can bound the above two terms separately. To bound (a), note that for any $1 \leq k \leq K$,

$$\begin{aligned}
&\Pr(S_k \geq kT \,|\, S_{k-1} \leq (k-1)T - 1) \\
&\quad = \Pr(d_h(X_{kT+1}, Y_{kT+1}) \geq \epsilon_0 \,|\, S_{k-1} \leq (k-1)T - 1) \\
&\quad \leq \frac{\mathbb{E}\left[d_h(X_{kT+1}, Y_{kT+1}) \,|\, S_{k-1} \leq (k-1)T - 1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right]}{\epsilon_0} \\
&\quad \leq \frac{c_1}{\epsilon_0} \mathbb{E}\left[d_v(X_{kT}, Y_{kT}) \,|\, S_{k-1} \leq (k-1)T - 1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right] \\
&\quad \leq \frac{c_1}{\epsilon_0} \mathbb{E}\left[(c_1 c_2)^{kT - S_{k-1} - 1} d_v(X_{S_{k-1}}, Y_{S_{k-1}}) \,|\, S_{k-1} \leq (k-1)T - 1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right] \\
&\quad \leq \frac{c_1 (c_1 c_2)^T M}{\epsilon_0}
\end{aligned}$$

To bound (b) we make use of the fact that at each random time S_k we have at least a $1 - \delta_0$ chance of setting $X_{S_k}(v) = Y_{S_k}(v)$. Therefore,

$$\Pr(X_{KT}(v) \neq Y_{KT}(v) \,|\, S_1 \leq T-1, \ldots S_K \leq KT - 1) \leq \delta_0^K.$$

Then for $K = \frac{\log(2/\delta)}{\log(1/\delta_0)}$ and $T = \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{2 c_1 K M}{\delta \epsilon_0}\right)$,

$$\Pr(X_{KT}(v) \neq Y_{KT}(v)) \leq \frac{c_1 (c_1 c_2)^T K M}{\epsilon_0} + \delta_0^K \leq \delta.$$

The lemma follows by our choice of K and T. □

Lemma C.1. *(a) There exists a coupling (X, Y) of $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ such that*

$$\mathbb{E}\left[(X-Y)^2\right] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2.$$

(b) There exists a coupling (X, Y) of $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ such that

$$\begin{aligned} \Pr(X \neq Y) &\leq \frac{|\mu_X - \mu_Y|}{2\sigma} \quad \textit{and} \\ \mathbb{E}\left[(X-Y)^2 \,|\, X \neq Y\right] &\leq 4\sigma^2 \left[1 + \frac{|\mu_X - \mu_Y|}{\sqrt{2\pi}\sigma} + \left(\frac{|\mu_X - \mu_Y|}{2\sigma}\right)^2\right]. \end{aligned}$$

Proof. Part (a) follows from a more general result [78]. To prove part (b), we introduce some notation. Let $\bar{\mu} = \frac{\mu_X + \mu_Y}{2}$. Assume w.l.o.g. $\bar{\mu} = 0$, $\mu_X = -\mu$, and $\mu_Y = \mu$ for some $\mu \geq 0$. Let f_X and f_Y denote the p.d.f.'s of $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively. Now define three more p.d.f.'s:

$$\begin{aligned} f_S(x) &= \frac{\min\left(f_X(x), f_Y(x)\right)}{Z_S} \qquad \text{for } x \in \mathbb{R} \\ f_U(x) &= \frac{f_Y(x) - f_X(x)}{Z_U} \qquad \text{for } x \geq 0 \\ f_L(x) &= \frac{f_X(x) - f_Y(x)}{Z_L} \qquad \text{for } x \leq 0 \end{aligned}$$

Here Z_S, Z_U, and Z_L are chosen so that their respective distributions integrate to 1. It is not too hard to work out that

$$\begin{aligned} Z_S &= 2\left(1 - \Phi\left(\frac{\mu}{\sigma}\right)\right) = 1 - \operatorname{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) \\ Z_U = Z_L &= \Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma}\right) = \operatorname{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) \end{aligned}$$

Here $\Phi(\cdot)$ denotes cumulative distribution function for the standard normal distribution and $\operatorname{erf}(\cdot)$ denotes the error function. Figure C.1 helps explain the picture.

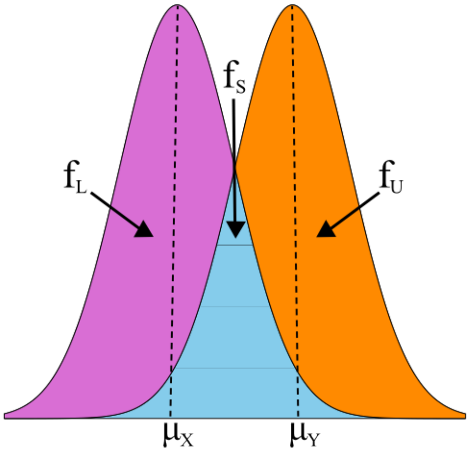


Figure C.1. Illustration of the unnormalized densities f_S, f_U, f_L.

Then our coupling is the following.

1. Draw $S \sim f_S$, $U \sim f_U$, $L = -U$.
2. With probability Z_S, set $X = S = Y$.
3. With probability $1 - Z_S$, set $X = L$ and $Y = U$.

It is not hard to see that (X, Y) is a valid coupling of f_X and f_Y. We now turn to the two claims of this coupling. The first is easy:

$$\Pr(X \neq Y) = 1 - Z_S = \operatorname{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) = \operatorname{erf}\left(\frac{|\mu_X - \mu_Y|}{2\sigma\sqrt{2}}\right) \leq \frac{|\mu_X - \mu_Y|}{2\sigma}.$$

Now we turn to the second claim. To handle this, we will first introduce two more random variables.

- Let U_X be distributed according to $f_{U_X}(x) = \frac{f_X(x)}{1-\Phi(\frac{\mu}{\sigma})}$.
- Let U_Y be distributed according to $f_{U_Y}(x) = \frac{f_Y(x)}{1-\Phi(-\frac{\mu}{\sigma})}$.

Then we can rewrite our objective to bound as

$$\mathbb{E}[(X-Y)^2 \,|\, X \neq Y] = \mathbb{E}[(U-L)^2] \;=\; 4\mathbb{E}[U^2]$$

$$= \frac{4}{Z_U}\left[\left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right)\mathbb{E}[U_Y^2] - \left(1 - \Phi\left(\frac{\mu}{\sigma}\right)\right)\mathbb{E}[U_X^2]\right]$$

It is easy to see that U_X and U_Y follow truncated normal distributions. Their moments can be worked out according to formulas given by [59]. In particular we have for $\epsilon = \frac{|\mu_X - \mu_Y|}{2\sigma}$

$$\begin{aligned}
\mathbb{E}[(X-Y)^2 \,|\, X \neq Y] &= \frac{4}{Z_U}\left[(\mu^2+\sigma^2)Z_U + \frac{2\mu}{\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right] \\
&= 4\sigma^2\left[1+\epsilon^2+\frac{2\epsilon}{\sqrt{2\pi}\mathrm{erf}(\epsilon/\sqrt{2})}\right] \\
&\leq 4\sigma^2\left[1+\epsilon^2+\sqrt{\frac{2}{\pi}}\left(\epsilon+\sqrt{\frac{\pi}{2}}\right)\right] \\
&= 4\sigma^2\left[2+\epsilon\sqrt{\frac{2}{\pi}}+\epsilon^2\right]
\end{aligned}$$

where the inequality in the third line comes from the inequality $\frac{x}{\mathrm{erf}(x/\sqrt{2})} \leq x + \sqrt{\frac{\pi}{2}}$. □

Lemma 6.10. *The following holds.*

(a) $P_{GG}^{(v)}$, $P_{GG}^{(h)}$, $P_{GN}^{(v)}$ *are* $\|W\|_F^2$*-contractive.*

(b) $P_{GN}^{(h)}$ *is* $\frac{5}{4}\|W\|_F^2$*-contractive.*

(c) $P_{GG}^{(v)}$ *and* $P_{GN}^{(v)}$ *are* $(\epsilon_0, \delta_0, M)$*-gamble admissible for* $\epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$, $\delta_0 = 1/4$, *and*

$$M \;=\; 4\|\sigma\|_2^2 + \sqrt{\frac{2}{\pi}}\frac{\left\|(W\sigma)^T\right\|_{2,1}}{\left\|(W/\sigma)^T\right\|_{2,1}} + \left(\frac{\|W\|_F}{2\|(W/\sigma)^T\|_{2,1}}\right)^2$$

where W/σ *and* $W\sigma$ *denote* $n\times m$ *matrices whose entries are* W_{ij}/σ_i *and* $W_{ij}\sigma_i$*, respectively*

Proof. To prove part (a), we need only show the result for $P_{GG}^{(v)}$; the bounds for $P_{GG}^{(h)}$ and $P_{GN}^{(v)}$ will follow symmetrically.

Recall that our distance is ℓ_2^2-distance $d_v(x,y) := \sum_{i=1}^n (x(v_i) - y(v_i))^2$. To see that $P_{GG}^{(v)}$ is contractive, let $x, y \in \Omega$ be given. We will construct our contractive coupling (X,Y) by coupling each visible node $X(v_i)$ independently. In particular, we will use the coupling from Lemma C.1(a) to couple together the marginal distributions $N(a_i+\sum_{j=1}^m W_{ij}x(h_j),\, \sigma_i^2)$ and $N(a_i + \sum_{j=1}^m W_{ij}y(h_j),\, \sigma_i^2)$. By Lemma C.1(a), we have

$$\begin{aligned}
\mathbb{E}[d_v(X,Y)] &= \sum_{i=1}^n \mathbb{E}\left[(X(v_i) - Y(v_i))^2\right] \\
&\leq \sum_{i=1}^n \left(\sum_{j=1}^m W_{ij}\left(x(h_j) - y(h_j)\right)\right)^2 \\
&\leq \left(\sum_{i=1}^n \sum_{j=1}^m W_{ij}^2\right) \sum_{j=1}^m \left(x(h_j) - y(h_j)\right)^2 \\
&= \|W\|_F^2 \, d_h\left(x, y\right).
\end{aligned}$$

To prove part (b), we will couple each unit h_j independently as follows.

1. Let (Z_j, Z_j') be the coupling from Lemma C.1(a) of

$$N\left(\sum_{i=1}^n W_{ij}x(v_i), \sigma\left(\sum_{i=1}^n W_{ij}x(v_i)\right)\right) \quad \text{and} \quad N\left(\sum_{i=1}^n W_{ij}y(v_i), \sigma\left(\sum_{i=1}^n W_{ij}y(v_i)\right)\right).$$

2. Let $X(h_j) = \max(0, Z_j)$ and $Y(h_j) = \max(0, Z_j')$.

Then by the definition of NReLU, X and Y have the correct marginal distributions. To see that they are contractive, note first that for each h_j,

$$\begin{aligned}
\mathbb{E}\left[(X(h_j) - Y(h_j))^2\right] &\leq \mathbb{E}\left[\left(Z_j - Z_j'\right)^2\right] \\
&= \left(\sum_{i=1}^n W_{ij}\left(x(v_i) - y(v_i)\right)\right)^2 + \left(\sigma\left(\sum_{i=1}^n W_{ij}x(v_i)\right) - \sigma\left(\sum_{i=1}^n W_{ij}y(v_i)\right)\right)^2
\end{aligned}$$

where the equality comes from Lemma C.1(a). Thus,

$$\mathbb{E}[d_h(X,Y)] \leq \sum_{j=1}^m \left(\sum_{i=1}^n W_{ij}\left(x(v_i) - y(v_i)\right)\right)^2$$

$$
\begin{aligned}
&+\sum_{j=1}^{m}\left(\sigma\left(\sum_{i=1}^{n} W_{ij}x(v_i)\right)-\sigma\left(\sum_{i=1}^{n} W_{ij}y(v_i)\right)\right)^2\\
&\leq \|W\|_F^2\, d_v\left(x,y\right)+\sum_{j=1}^{m}\left(\frac{1}{2}\sum_{i=1}^{n} W_{ij}(x(v_i)-y(v_i))\right)^2\\
&\leq \|W\|_F^2\, d_v\left(x,y\right)+\frac{1}{4}\|W\|_F^2\, d_v\left(x,y\right)\\
&=\frac{5}{4}\|W\|_F^2\, d_v\left(x,y\right).
\end{aligned}
$$

To prove part (c), it will suffice to prove that $P_{GG}^{(v)}$ is gamble admissible; the gamble admissibility of $P_{GN}^{(v)}$ will follow by symmetry. To do this, we will construct a gamble coupling (X, Y) by independently coupling the visible nodes v_i according to the coupling from Lemma C.1(b). The probability that we set $X(v) \neq Y(v)$ is bounded as

$$
\begin{aligned}
\Pr(X(v)\neq Y(v)) &= Pr\left(\exists v_i \text{ s.t. } X(v_i)\neq Y(v_i)\right)\\
&\leq \sum_{i=1}^{n}\Pr(X(v_i)\neq Y(v_i))\\
&\leq \sum_{i=1}^{n}\frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j)-y(h_j)\right)\right|}{2\sigma_i}\\
&\leq \frac{1}{2}\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}\left(\frac{W_{ij}}{\sigma_i}\right)^2}\sqrt{\sum_{j=1}^{m}\left(x(h_j)-y(h_j)\right)^2}\\
&=\frac{\left\|(W/\sigma)^T\right\|_{2,1}}{2}\sqrt{d_h\left(x,y\right)}
\end{aligned}
$$

Similarly we can bound $\mathbb{E}[d_v(X,Y)\,|\,X(v)\neq Y(v)]$ by

$$
\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\left[(X(v_i)-Y(v_i))^2\,|\,X(v_i)\neq Y(v_i)\right]\\
&\quad\leq \sum_{i=1}^{n}4\sigma_i^2\left[1+\frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j)-y(h_j)\right)\right|}{\sqrt{2\pi}\sigma_i}+\left(\frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j)-y(h_j)\right)\right|}{2\sigma_i}\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} 4\sigma_i^2 + 2\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \sigma_i W_{ij} \left(x(h_j) - y(h_j)\right) \right| + \sum_{i=1}^{n} \left(\sum_{j=1}^{m} W_{ij} \left(x(h_j) - y(h_j)\right) \right)^2 \\
&\leq \sum_{i=1}^{n} 4\sigma_i^2 + 2\sqrt{\frac{2}{\pi}} \left\|(W\sigma)^T\right\|_{2,1} \sqrt{d_h\left(x, y\right)} + \|W\|_F^2 \, d_h\left(x, y\right).
\end{aligned}
$$

Plugging in $d_h(x, y) \leq \epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$ finishes the proof.

□

C.3 Proofs from Section 6.5

Theorem 6.12. *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there is a weight matrix $W \in \mathbb{R}^{n \times m}$ satisfying*

$$
\|W\|_{max} \;\leq\; \frac{2}{\min(n, m)} \ln\left(4T(n+m)\right)
$$

such that the Gibbs sampler over the RBM with zero bias and weight matrix W has mixing rate bounded as $\tau_{mix} \geq T$.

Proof. Let $r = \frac{2}{\min(n,m)} \ln\left(4T(n+m)\right)$. Choose a canonical configuration x such that exactly half of the $x(v_i)$'s are 1 and exactly half of the $x(h_j)$'s are 1. Now let $W \in \mathbb{R}^{n \times m}$ such that $W_{ij} = r$ if $x(v_i) = x(h_j)$ and $-r$ otherwise. Let $\pi(\cdot)$ denote the Gibbs distribution for the RBM with weight matrix W and zero bias and let $S = \{x\}$ be the singleton set containing only the canonical configuration. Note that if $\bar{x}$ satisfies that $\bar{x}(v_i) = 1$ iff $x(v_i) = 0$ and $\bar{x}(h_j) = 1$ iff $x(h_j) = 0$, then $\pi(x) = \pi(\bar{x})$. Thus, $\pi(S) \leq 1/2$.

It is not hard to see $\Pr(X(h_j) \neq x(h_j) \,|\, x(v)) = \sigma\left(-\frac{nr}{2}\right)$ for all $j \in [m]$, where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid as before. Similarly, for any $i \in [n]$, $\Pr(X(v_i) \neq x(v_i) \,|\, x(h)) = \sigma\left(-\frac{mr}{2}\right)$. Thus,

$$
\Pr(\text{leave state } x) \leq \frac{m}{1 + \exp\left(\frac{nr}{2}\right)} + \frac{n}{1 + \exp\left(\frac{mr}{2}\right)} \leq \frac{1}{4T}
$$

Thus the conductance of S (and therefore Φ^*) is upper bounded as

$$\Phi(S) = \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x)\mathrm{Pr}(\text{we transition from } x \text{ to } y) = \mathrm{Pr}(\text{leave state } x) \leq \frac{1}{4T}$$

Theorem 4.2 completes the proof. □

Lemma C.2. $\Phi(x) \leq 1 - \sqrt{1 - \exp\left(-\frac{x^2}{2}\right)}$ *for* $x \leq 0$.

Proof. We begin by writing $\Phi(\cdot)$ in terms of the error function:

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right).$$

Thus it suffices to prove $\mathrm{erf}\,(x)^2 \geq 1 - e^{-x^2}$. By calculus, we have

$$\begin{aligned}
\mathrm{erf}(x)^2 &= \frac{4}{\pi} \int_0^x \int_0^x e^{-(s^2+t^2)}\, ds\, dt \\
&\geq \frac{4}{\pi} \int_0^{\pi/2} \int_0^x r e^{-r^2}\, dr\, d\theta \\
&= \frac{4}{\pi} \int_0^{\pi/2} \left[-\frac{1}{2} e^{-r^2} \Big|_{r=0}^{x} \right] d\theta \\
&= \frac{4}{\pi} \int_0^{\pi/2} \frac{1}{2}\left(1 - e^{-x^2}\right) d\theta \\
&= 1 - e^{-x^2}
\end{aligned}$$

where the inequality comes from the fact that $e^{-(s^2+t^2)} \geq 0$ and the quarter circle of radius x centered at the origin and lying in the first quadrant is a subset of the square $[0, x]^2$. □

Theorem 6.13. *Let* $T, B > 0$ *and* $n, m \in \mathbb{N}$ *be even positive integers. Then there exists weight matrix* $W \in \mathbb{R}^{n \times m}$ *s.t.*

$$\|W\|_{max} \;\leq\; \frac{1}{\min(n,m)}\left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$$

such that the B*-truncated chain of the Gibbs sampler for the Gaussian-Gaussian RBM* *with no biases and unit variances mixes in time $\tau_{mix} \geq T$.*

Proof. Let $r = \frac{1}{\min(n,m)}\left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$. Let $\mathcal{I}_-, \mathcal{I}_+$ be an even partition of $[n]$, i.e. $|\mathcal{I}_-| = n/2 = |\mathcal{I}_+|$. Similarly, let $\mathcal{J}_-, \mathcal{J}_+$ be an even partition of $[m]$. Define

$$W_{ij} = \begin{cases} r & \text{if } (i,j) \in \mathcal{I}_- \times \mathcal{J}_- \cup \mathcal{I}_+ \times \mathcal{J}_+ \\ -r & \text{else} \end{cases}$$

$$S_v = \{x(v) \in [-B,B]^n \,:\, x(v_i) \geq B/2 \text{ if } i \in \mathcal{I}_+ \text{ and } x(v_i) \leq -B/2 \text{ else}\}$$

$$S_h = \{x(h) \in [-B,B]^m \,:\, x(h_j) \geq B/2 \text{ if } j \in \mathcal{J}_+ \text{ and } x(h_j) \leq -B/2 \text{ else}\}$$

Then our low conductance set of configurations is $S = S_v \times S_h$. Note that the c.d.f.'s of the conditional distributions for the B-thresholded chain are exactly the same as the regular normal distribution for points within $[-B,B]$. That is, given $x \in \Omega$ and $p \in (-B,B)$, for any hidden node h_j and visible node v_i

$$\begin{aligned} P(X(h_j) < p \,|\, x(v)) &= \Phi\left(p - \sum_{i=1}^{n} W_{ij}\, x(v_i)\right) \\ P(X(v_i) < p \,|\, x(h)) &= \Phi\left(p - \sum_{j=1}^{m} W_{ij}\, x(h_j)\right) \end{aligned}$$

For $x \in S$ and $j \in \mathcal{J}_+$, we have by Lemma C.2,

$$\begin{aligned} P(X(h_j) < B/2 \,|\, x(v)) &= \Phi\left(\frac{B}{2} - r\left(\sum_{i\in\mathcal{I}+} x(v_i) - \sum_{i\in\mathcal{I}_-} x(v_i)\right)\right) \\ &\leq \Phi\left(\frac{B}{2}(1 - rn)\right) \\ &\leq 1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1-rn)^2\right)}. \end{aligned}$$

Symmetric inequalities also hold for $P(X(h_j) > -B/2 \,|\, x(v))$ when $j \in \mathcal{J}_-$. Additionally, for $i \in \mathcal{I}_+$ and $i' \in \mathcal{I}_-$,

$$P(X(v_i) < B/2 \,|\, x(h)), P(X(v_{i'}) > -B/2 \,|\, x(h)) \leq 1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1-rm)^2\right)}.$$

Therefore, given that the current state of our chain Y_t is in S, we can bound the probability that we transition out of S in the next step as

$$\begin{aligned} P(Y_{t+1} \notin S \,|\, Y_t \in S) \leq m &\left(1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1-rn)^2\right)}\right) \\ &+ n\left(1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1-rm)^2\right)}\right). \end{aligned}$$

Plugging in our value for r gives us an upper bound of $\frac{1}{4T}$. Theorem 4.2 completes the proof. □

C.4 Proofs from Section 6.6

[61, 69, 45] technically deal with Ising (or spin glass) models as opposed to Boltzmann machines. As the following lemma demonstrates, however, the partition functions of these models differs only by an easily computable constant. Thus, they are approximation-preserving interreducible in the sense of [39].

Lemma C.3. *Let $G = (V, E)$ be a graph, $W_{ij} \in \mathbb{R}$ for all $(i,j) \in E$, $b_i \in \mathbb{R}$ for all $i \in V$, and define*

$$Z_{Ising}(G, W, b) = \sum_{x \,:\, V \to \{-1,1\}^V} \exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i \in V} b_i x(i)\right)$$

as the Ising partition function and

$$Z_{Boltzmann}(G,W,b) = \sum_{x\,:\,V\to\{0,1\}^V} \exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i\in V} b_i x(i)\right)$$

as the Boltzmann partition function then $CZ_{Ising}(G,W,b) = Z_{Boltzmann}(G,W',b')$ *where*

$$\begin{aligned} W' &= 4W \\ b'_i &= 2b_i - 2\sum_{j\ s.t.(i,j)\in E} W_{ij} \\ C &= \exp\left(\sum_{i\in V} b_i - \sum_{(i,j)\in E} W_{ij}\right) \end{aligned}$$

Proof. The key idea is to identify every Ising configuration $x : V \to \{-1,1\}^V$ with a Boltzmann configurations $y : V \to \{0,1\}^V$. The convention we will take is $y(i) = \frac{1}{2}(x(i)+1)$, which has the effect of identifying the spin -1 with 0 and 1 with 1. Then for any Ising/Boltzmann corresponding pair x, y, we have

$$\begin{aligned} &\exp\left(\sum_{(i,j)\in E} W'_{ij}y(i)y(j) + \sum_{i\in V} b'_i y(i)\right) \\ &= \exp\left(\sum_{(i,j)\in E} 4W_{ij}y(i)y(j) + \sum_{i\in V} y(i)\left(2b_i - 2\sum_{j\text{ s.t.}(i,j)\in E} W_{ij}\right)\right) \\ &= \exp\left(\sum_{(i,j)\in E} W_{ij}(x(i)+1)(x(j)+1) + \sum_{i\in V}(x(i)+1)\left(b_i - \sum_{j\text{ s.t.}(i,j)\in E} W_{ij}\right)\right) \\ &= \exp\left(\sum_{(i,j)\in E} W_{ij}(x(i)+x(j)+1) - \sum_{i\in V}(x(i)+1)\left(\sum_{j\text{ s.t.}(i,j)\in E} W_{ij}\right) + \sum_{i\in V} b_i\right)\cdot \\ &\quad \exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i\in V} b_i x(i)\right) \end{aligned}$$

$$= C \exp\left(\sum_{(i,j)\in E} W_{ij} x(i) x(j) + \sum_{i\in V} b_i x(i)\right).$$

Because the mapping from Ising to Boltzmann configurations is bijective, it then holds that

$$\begin{aligned} Z_{\text{Boltzmann}}(G, W', b') &= \sum_{y\,:\,V\to\{0,1\}^V} \exp\left(\sum_{(i,j)\in E} W'_{ij} y(i) y(j) + \sum_{i\in V} b'_i y(i)\right) \\ &= \sum_{x\,:\,V\to\{-1,1\}^V} C \exp\left(\sum_{(i,j)\in E} W_{ij} x(i) x(j) + \sum_{i\in V} b_i x(i)\right) \\ &= C Z_{\text{Ising}}(G, W, b). \qquad \square \end{aligned}$$

Appendix D

Supplementary material for Chapter 7

D.1 Remark from Section 7.3

In Section 7.3, the remark after the definition of average splitting stated that there exist hypothesis classes V for which there are many points which 1/4-split E for any $E \subset \binom{V}{2}$ but for which any $x \in \mathcal{X}$ satisfies

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \approx \Phi(V).$$

Here we formally prove this statement.

Consider the hypothesis class of homogeneous linear separators and let $V = \{e_1, \ldots, e_n\} \subset \mathcal{H}$ where e_k is the k-th unit coordinate vector. Let the data distribution be uniform over the n-sphere and the prior distribution π be uniform over V. As a subset of the homogeneous linear separators, V has splitting index $(1/4, \epsilon, \Theta(\epsilon))$ [31, Theorem 10].

On the other hand, for any $i \neq j$, $d(h_i, h_j) = 1/2$. This implies that

$$\Phi(V) \;=\; \Pr(h \neq h')\mathbb{E}_{h,h'}[d(h,h') \,|\, h \neq h'] \;=\; \frac{n-1}{2n}.$$

Moreover, any query $x \in \mathcal{X}$ eliminates at most half the hypotheses in V in the worst case.

Therefore, for all $x \in \mathcal{X}$,

$$\max\{\Phi(V_x^+), \Phi(V_x^-)\} \;\geq\; \frac{(n/2-1)}{2(n/2)} \;=\; \left(\frac{n-2}{n-1}\right)\Phi(V).$$

D.2 Proofs of Lemma 7.5 and Lemma 7.7

The proofs in this section rely crucially on two concentration inequalities. The first is known as Hoeffding's inequality [55].

Lemma D.1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$ and let $X = \sum X_i$ and $\mu = \mathbb{E}[X]$. Then for $t > 0$,*

$$\Pr(X - \mu \geq t) \leq \exp\left(-\frac{2t^2}{n}\right)$$

Our other tool will be the following multiplicative Chernoff-Hoeffding bound [3].

Lemma D.2. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$ and let $X = \sum X_i$ and $\mu = \mathbb{E}[X]$. Then for $0 < \beta < 1$,*

(i) $\Pr(X \leq (1-\beta)\mu) \leq \exp\left(-\frac{\beta^2\mu}{2}\right)$ *and*

(ii) $\Pr(X \geq (1+\beta)\mu) \leq \exp\left(-\frac{\beta^2\mu}{3}\right)$.

We now turn to the proof of Lemma 7.5.

Lemma 7.5. *Let $\rho, \epsilon, \delta_0 > 0$ be given. Suppose that version space V satisfies $\Phi(V) > \epsilon$. In* SELECT*, fix a round t and data point $x \in \mathcal{X}$ that exactly ρ-average splits V (that is, $\max\{\pi|_V(V_x^+)^2\Phi(V_x^+),\ \pi|_V(V_x^-)^2\Phi(V_x^-)\} = (1-\rho)\Phi(V)$). If*

$$m_t \geq \frac{48}{\widehat{\rho}_t^2\epsilon}\log\frac{4}{\delta_0} \qquad \textit{and} \qquad n_t \geq \max\left\{\frac{32}{\widehat{\rho}_t^2\widehat{\Phi}_t}, \frac{40}{\widehat{\Phi}_t^2}\right\}\log\frac{4}{\delta_0}$$

then with probability $1 - \delta_0$,

(a) $\widehat{\Phi}_t \geq (1 - \widehat{\rho}_t/4)\Phi(V)$*;*

(b) if $\rho \leq \widehat{\rho}_t/2$*, then* $\frac{1}{n_t} \max\{\psi(E_x^+), \psi(E_x^-)\} > (1 - \widehat{\rho}_t)\widehat{\Phi}_t$*; and*

(c) if $\rho \geq 2\widehat{\rho}_t$*, then* $\frac{1}{n_t} \max\{\psi(E_x^+), \psi(E_x^-)\} \leq (1 - \widehat{\rho}_t)\widehat{\Phi}_t$*.*

Proof. In round t, let $\widehat{\rho} := \widehat{\rho}_t$, $\widehat{\Phi} := \widehat{\Phi}_t$, $m := m_t$, and $n := n_t$. For (a), recall $\widehat{\Phi} = \frac{1}{m}\psi(E')$ for $E' \sim (\pi|_V)^{2\times m}$. By Lemma D.2, we have for $\beta_0 > 0$

$$\Pr\Big((1-\beta_0)\Phi(V) \leq \widehat{\Phi} \leq (1+\beta_0)\Phi(V)\Big) \geq 1 - 2\exp\left(-\frac{m\beta_0^2\epsilon}{3}\right).$$

Taking $m \geq \frac{3}{\beta_0^2\epsilon}\log\left(\frac{4}{\delta_0}\right)$, we have the above probability is at least $1 - \delta_0/2$. Let us condition on this event occurring.

To see (b), say w.l.o.g. $\left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \Phi(V_x^+) = (1-\rho)\Phi(V)$. Then,

$$\Pr\left(\frac{1}{n}\psi(E_x^+) \leq (1-\widehat{\rho})\widehat{\Phi}\right) \leq \Pr\left(\frac{1}{n}\psi(E_x^+) \leq (1-\widehat{\rho})(1+\beta_0)\Phi(V)\right).$$

Let β satisfy $(1-\beta)(1-\rho) = (1-\widehat{\rho})(1+\beta_0)$. By Lemma D.2 (i),

$$\begin{aligned}
\Pr\left(\frac{1}{n}\psi(E_x^+) \leq (1-\widehat{\rho})\widehat{\Phi}\right) &\leq \Pr\left(\frac{1}{n}\psi(E_x^+) \leq (1-\beta)(1-\rho)\Phi(V)\right) \\
&\leq \exp\left(-\frac{n\beta^2(1-\rho)\Phi(V)}{2}\right) \\
&\leq \exp\left(-\frac{n(1-\rho)\widehat{\Phi}}{2(1+\beta_0)} \cdot \left[1 - \frac{(1-\widehat{\rho})(1+\beta_0)}{1-\rho}\right]^2\right) \\
&\leq \exp\left(-\frac{n(1-\widehat{\rho}/2)\widehat{\Phi}}{2(1+\beta_0)} \cdot \left[1 - \frac{(1-\widehat{\rho})(1+\beta_0)}{1-\widehat{\rho}/2}\right]^2\right).
\end{aligned}$$

Taking $\beta_0 \leq \widehat{\rho}/4$, the above is less than $\exp\left(-\frac{n\widehat{\Phi}\widehat{\rho}^2}{32}\right)$. With n as in the lemma statement and combined with our results on the concentration of $\widehat{\Phi}$, we have with probability $1 - \delta_0$,

$$\frac{1}{n}\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} > (1-\widehat{\rho})\widehat{\Phi}.$$

To see (c), suppose now that w.l.o.g. $\left(\frac{\pi(V_x^-)}{\pi(V)}\right)^2 \Phi(V_x^-) \leq \left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \Phi(V_x^+) = (1-\rho)\Phi(V)$. We need to consider two cases.

Case 1: $\rho \leq 1/2$. Taking β such that $(1+\beta)(1-\rho) = (1-\widehat{\rho})(1-\beta_0)$, we have by Lemma D.2 (ii),

$$\begin{aligned}
\Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\widehat{\rho})\widehat{\Phi}\right) &\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\widehat{\rho})(1-\beta_0)\Phi(V)\right) \\
&= \Pr\left(\frac{1}{n}\psi(E_x^+) > (1+\beta)(1-\rho)\Phi(V)\right) \\
&\leq \exp\left(-\frac{n\beta^2(1-\rho)\Phi(V)}{3}\right) \\
&\leq \exp\left(-\frac{n(1-\rho)\widehat{\Phi}}{3(1+\beta_0)} \cdot \left[\frac{(1-\widehat{\rho})(1-\beta_0)}{1-\rho} - 1\right]^2\right) \\
&\leq \exp\left(-\frac{n\widehat{\Phi}}{6(1+\beta_0)} \cdot \left[\frac{(1-\widehat{\rho})(1-\beta_0)}{1-2\widehat{\rho}} - 1\right]^2\right).
\end{aligned}$$

Taking $\beta_0 \leq \widehat{\rho}/4$, the above is less than $\exp\left(-\frac{n\widehat{\Phi}\widehat{\rho}^2}{12}\right)$. Note this also implies

$$\Pr\left(\frac{1}{n}\psi(E_x^-) > (1-\widehat{\rho})\widehat{\Phi}\right) \leq \exp\left(-\frac{n\widehat{\Phi}\widehat{\rho}^2}{12}\right)$$

since $\left(\frac{\pi(V_x^-)}{\pi(V)}\right)^2 \Phi(V_x^-) \leq \left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \Phi(V_x^+)$.

Case 2: $\rho > 1/2$. Taking $\beta_0 \leq 1/16$, we have

$$\begin{aligned}
\Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\widehat{\rho})\widehat{\Phi}\right) &\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\widehat{\rho})(1-\beta_0)\Phi(V)\right) \\
&\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\rho)\Phi(V) + (\rho-\widehat{\rho}-\beta_0)\Phi(V)\right) \\
&\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\rho)\Phi(V) + \left(\frac{\rho}{2}-\beta_0\right)\Phi(V)\right) \\
&\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\rho)\Phi(V) + \left(\frac{1}{4}-\beta_0\right)\Phi(V)\right) \\
&\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\rho)\Phi(V) + \frac{\frac{1}{4}-\beta_0}{1+\beta_0}\widehat{\Phi}\right)
\end{aligned}$$

$$\leq \Pr\left(\frac{1}{n}\psi(E_x^+) > (1-\rho)\Phi(V) + \frac{3}{17}\widehat{\Phi}\right)$$

By Lemma D.1, the above is less than $\exp\left(-\frac{n\widehat{\Phi}^2}{40}\right)$. Note this also implies

$$\Pr\left(\frac{1}{n}\psi(E_x^-) > (1-\widehat{\rho})\widehat{\Phi}\right) \leq \exp\left(-\frac{n\widehat{\Phi}^2}{40}\right)$$

since $\left(\frac{\pi(V_x^-)}{\pi(V)}\right)^2 \Phi(V_x^-) \leq \left(\frac{\pi(V_x^+)}{\pi(V)}\right)^2 \Phi(V_x^+)$. Regardless of which case we are in, we have for n as in the lemma statement, with probability $1-\delta_0$,

$$\frac{1}{n}\max\left\{\psi(E_x^+), \psi(E_x^-)\right\} \leq (1-\widehat{\rho})\widehat{\Phi}. \qquad \square$$

We next provide the proof of Lemma 7.7.

Lemma 7.7. *The following holds for DBAL:*

(a) Suppose that for all $t = 1, 2, \ldots, K$ that $\Phi(V_t) > \epsilon$. Then the probability that the termination condition is ever true for any of those rounds is bounded above by $K\exp\left(-\frac{\epsilon n}{32}\right)$.

(b) Suppose that for some $t = 1, 2, \ldots, K$ that $\Phi(V_t) \leq \epsilon/2$. Then the probability that the termination condition is not true in that round is bounded above by $K\exp\left(-\frac{\epsilon n}{48}\right)$.

Proof. Recall that the termination condition from DBAL is $\frac{1}{n}\psi(E) < \frac{3\epsilon}{4}$ for $E \sim (\pi|_V)^{2\times n}$.

Part (a) follows from plugging in $\beta = \frac{1}{4}$ into Lemma D.2 (i) and taking a union bound over rounds $1, \ldots, K$.

Similarly, part (b) follows from plugging in $\beta = \frac{1}{4}$ into Lemma D.2 (ii) and taking a union bound over rounds $1, \ldots, K$. □

Appendix E

Supplementary material for Chapter 8

E.1 Proof of Lemma 8.2

Lemma 8.2. *Suppose $\pi = N(0, \sigma_o^2 I_d)$, $\ell(\cdot,\cdot)$ is the squared-loss, and we have observed $(x_1, y_1), \cdots, (x_t, y_t)$. If $\Phi \in \mathbb{R}^{t\times d}$ denotes the matrix*

$$\Phi = \begin{bmatrix} — & \phi(x_1) & — \\ — & \phi(x_2) & — \\ & \vdots & \\ — & \phi(x_t) & — \end{bmatrix}.$$

then π_t is $N(\widehat{\mu}, \widehat{\Sigma})$ where $\widehat{\Sigma} = \left(2\beta\Phi^T\Phi + \frac{1}{\sigma_o^2} I_d\right)^{-1}$ and $\widehat{\mu} = 2\beta\widehat{\Sigma}\Phi^T y$.

Proof. By expanding the form of $N(\widehat{\mu}, \widehat{\Sigma})$, we first see

$$\begin{aligned} N(w \,|\, \widehat{\mu}, \widehat{\Sigma}) &\propto \exp\left(-\frac{1}{2}(w-\widehat{\mu})^T\widehat{\Sigma}^{-1}(w-\widehat{\mu})\right) \\ &= \exp\left(-\frac{1}{2}w^T\widehat{\Sigma}^{-1}w - \frac{1}{2}\widehat{\mu}^T\widehat{\Sigma}^{-1}\widehat{\mu} + w^T\widehat{\Sigma}^{-1}\widehat{\mu}\right) \\ &\propto \exp\left(-\frac{1}{2}w^T\widehat{\Sigma}^{-1}w + w^T\widehat{\Sigma}^{-1}(2\beta\widehat{\Sigma}\Phi^T y)\right) \\ &= \exp\left(-\frac{1}{2}w^T(2\beta\Phi^T\Phi + \frac{1}{\sigma_o^2}I)w + 2\beta w^T\Phi^T y\right) \end{aligned}$$

$$= \exp\left(-\beta\left(w^T\Phi^T\Phi w - 2w^T\Phi^T y\right) - \frac{\|w\|^2}{2\sigma_o^2}\right)$$

On the other hand, we have

$$\begin{aligned}
\pi_t(w) &\propto \exp\left(-\beta\|y-\Phi w\|^2 - \frac{\|w\|^2}{2\sigma_o^2}\right) \\
&= \exp\left(-\beta\left(y^Ty + (\Phi w)^T\Phi w - 2(\Phi w)^T y\right) - \frac{\|w\|^2}{2\sigma_o^2}\right) \\
&\propto \exp\left(-\beta\left(w^T\Phi^T\Phi w - 2w^T\Phi^T y\right) - \frac{\|w\|^2}{2\sigma_o^2}\right)
\end{aligned}$$

Thus $\pi_t = N(\widehat{\mu}, \widehat{\Sigma})$. □

E.2 Proof of Lemma 8.6

Lemma E.1. *There exists a constant $c > 0$ such that the random variable $U_t = u(q_t; \pi_{t-1})$ satisfies*

$$\mathbb{E}[U_t \,|\, \mathcal{F}_{t-1}] \geq c\,\pi_{t-1}(g^*)(1-\pi_{t-1}(g^*))$$

for any round t where q_t is the query shown to the user at time t and the expectation is taken over the randomness of structural QBC.

Proof. Note that by the structural QBC strategy, the probability that $q \in \mathcal{Q}$ is selected is proportional to $\nu(q)\,u(q;\pi_{t-1})$. This implies

$$\mathbb{E}[U_t \,|\, \mathcal{F}_{t-1}] = \frac{\mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})^2]}{\mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})]} \geq \mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})].$$

Now take $\nu_o = \min_{q\in\mathcal{Q}} \nu(q)$ and $A_o = \max_{q\in\mathcal{Q}} |A(q)|$. Then for any $g \in \mathcal{G}$ s.t. $g \neq g^*$, there exists some $a \in \mathcal{A}$ such that $g(a) \neq g^*(a)$ which implies

$$\mathbb{E}_{q\sim\nu}[d(g,g^*;q)] = \sum_{q\in\mathcal{Q}} \frac{\nu(q)}{|A(q)|} \sum_{a\in A(q)} \mathbf{1}[g(a) \neq g^*(a)] \geq \frac{\nu_o}{A_o}.$$

Then we have

$$\begin{aligned}
\mathbb{E}[U_t \,|\, \mathcal{F}_{t-1}] &\geq \mathbb{E}_{q\sim\nu}[u(q;\pi_{t-1})] \\
&= \mathbb{E}_{g,g'\sim\pi_{t-1}}\left[\mathbb{E}_{q\sim\nu}[d(g,g';q)]\right] \\
&= \sum_{g,g'} \pi_{t-1}(g)\pi_{t-1}(g')\left[\mathbb{E}_{q\sim\nu}[d(g,g';q)]\right] \\
&\geq \pi_{t-1}(g^*) \sum_{g\neq g^*} \pi_{t-1}(g)\mathbb{E}_{q\sim\nu}[d(g,g^*;q)] \\
&\geq \frac{\nu_o}{A_o}\pi_{t-1}(g^*)(1-\pi_{t-1}(g^*)).
\end{aligned}$$ □

Lemma 8.6. *Suppose that $\mathcal{G}$ is finite and the user's feedback obeys Assumption 8.2. Then there exists a constant $c > 0$ such that for every round t*

$$\mathbb{E}[1-\gamma_t \,|\, \mathcal{F}_{t-1}] \;\geq\; c\,\pi_{t-1}(g^*)^2(1-\pi_{t-1}(g^*))^2$$

where $\gamma_t = \pi_{t-1}(\{g \in \mathcal{G} : g(a_t) = g^*(a_t)\})$, a_t is the atom the user provides feedback on, and the expectation is taken over the randomness in structural QBC and the user's response.

Proof. Suppose that structural QBC shows the user query q_t. Let $\epsilon > 0$ and $U_t = u(q_t;\pi_{t-1})$ be the random variable denoting the uncertainty of q_t. For an atom $a \in A(q_t)$, we say that a is *known* if

$$\pi_{t-1}(\{g : g(a) \neq g^*(a)\}) < \epsilon\, U_t.$$

And we say that a is *unknown* otherwise. By a union bound, we have

$$\pi_{t-1}(\{g : g(a) \neq g^*(a) \text{ for some known } a \in A(q_t)\}) \leq \epsilon\, U_t|A(q_t)|.$$

By Lemma 8.1, we have that for any $a \in \mathcal{A}$

$$\pi_{t-1}(\{g : g(a) \neq g^*(a)\}) \geq \frac{1}{2}u(a;\pi_{t-1}).$$

Moreover, there is *some* unknown atom $a^* \in A(q_t)$ with $u(a^*; \pi_{t-1}) \geq U_t$, implying

$$\pi_{t-1}(\{g : g(a^*) \neq g^*(a^*)\}) \geq \frac{1}{2} u(a^*; \pi_{t-1}) \geq \frac{U_t}{2}.$$

So with probability at least $U_t \left(\frac{1}{2} - \epsilon |A(q_t)|\right)$, a random draw from π_{t-1} gets all the known atoms correct and some unknown atom incorrect. Then conditioned on this event occurring, the user has probability at least p_o of correcting some unknown atom. So taking $\epsilon = \frac{1}{4A_o}$ where $A_o = \max |A(q)|$, we have

$$\begin{aligned}
\mathbb{E}\left[1 - \gamma_t \,|\, \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\pi_{t-1}(\{g : g(a_t) \neq g^*(a_t)\}) \,|\, \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\pi_{t-1}(\{g : g(a_t) \neq g^*(a_t)\}) | u(q_t; \pi_{t-1}) = U_t\right] \,|\, \mathcal{F}_{t-1}\right] \\
&\geq \mathbb{E}\left[\frac{U_t^2 p_o}{16 A_o} \,\middle|\, \mathcal{F}_{t-1}\right] \geq c_o \mathbb{E}[U_t^2 \,|\, \mathcal{F}_{t-1}] \geq c_o \mathbb{E}[U_t \,|\, \mathcal{F}_{t-1}]^2
\end{aligned}$$

where $c_o = p_o/(16A_o)$ is a positive constant. But by Lemma E.1, we know $\mathbb{E}[U_t] \geq c\, \pi_{t-1}(g^*)(1 - \pi_{t-1}(g^*))$ for some constant $c > 0$. The lemma follows by substitution. □

E.3 Proof of Lemma 8.10

Lemma E.2. *Suppose $\mathcal{G}$ is finite. For each round t, the query q_t under the structural QBC strategy satisfies* $\mathbb{E}[\mathrm{var}(q_t; \pi_{t-1}) \,|\, \mathcal{F}_{t-1}] \geq c\, \pi_{t-1}(g^*)(1 - \pi_{t-1}(g^*))$.

Proof. For a fixed query q, the probability that q gets chosen in round t can be written as

$$\Pr(\text{query } q \,|\, \mathcal{F}_{t-1}) = \frac{\nu(q)\, \mathrm{var}(q; \pi_{t-1})}{\sum_{q' \in \mathcal{Q}} \nu(q')\, \mathrm{var}(q'; \pi_{t-1})} = \frac{\nu(q)\, \mathrm{var}(q; \pi_{t-1})}{\mathbb{E}_{q' \sim \nu}[\mathrm{var}(q'; \pi_{t-1})]}.$$

Taking the expectation of $\mathrm{var}(q; \pi_{t-1})$ over this distribution gives us

$$\mathbb{E}[\mathrm{var}(q_t; \pi_{t-1}) \,|\, \mathcal{F}_{t-1}] \geq \frac{\mathbb{E}_{q \sim \nu}[\mathrm{var}(q; \pi_{t-1})^2]}{\mathbb{E}_{q \sim \nu}[\mathrm{var}(q; \pi_{t-1})]} \geq \mathbb{E}_{q \sim \nu}[\mathrm{var}(q; \pi_{t-1})].$$

Now take $\nu_o = \min_{q \in \mathcal{Q}} \nu(q)$ and

$$d_o = \min_{g \neq g^*} \max_{q \in \mathcal{Q}, a \in A(q)} \frac{\|g(a) - g^*(a)\|^2}{|A(q)|}.$$

For any particular $a \in \mathcal{A}$,

$$\begin{aligned} \mathrm{var}(a; \pi_{t-1}) &= \frac{1}{2} \sum_{g, g' \in \mathcal{G}} \pi_{t-1}(g)\, \pi_{t-1}(g')\, \|g(a) - g'(a)\|^2 \\ &\geq \pi_{t-1}(g^*) \sum_{g \neq g^*} \pi_{t-1}(g)\, \|g(a) - g^*(a)\|^2 \end{aligned}$$

Which implies

$$\begin{aligned} \mathbb{E}_{q \sim \nu}\left[\mathrm{var}(q; \pi_{t-1})\right] &= \sum_{q \in \mathcal{Q}} \frac{\nu(q)}{|A(q)|} \sum_{a \in A(q)} \mathrm{var}(a; \pi_{t-1}) \\ &\geq \sum_{q \in \mathcal{Q}} \frac{\nu(q)}{|A(q)|} \sum_{a \in A(q)} \pi_{t-1}(g^*) \sum_{g \neq g^*} \pi_{t-1}(g) \|g(a) - g^*(a)\|^2 \\ &= \pi_{t-1}(g^*) \sum_{g \neq g^*} \pi_{t-1}(g) \sum_{q \in \mathcal{Q}} \frac{\nu(q)}{|A(q)|} \sum_{a \in A(q)} \|g(a) - g^*(a)\|^2 \\ &\geq \nu_o d_o \pi_{t-1}(g^*)(1 - \pi_{t-1}(g^*)) \end{aligned}$$

□

Lemma 8.10. *Suppose $\mathcal{G}$ is finite and Assumption 8.2 holds. Then there exists a constant $c > 0$ such that for any round t*

$$\mathbb{E}[\mathrm{var}(a_t; \pi_{t-1}) \,|\, \mathcal{F}_{t-1}] \geq c\, \pi_{t-1}(g^*)^3 (1 - \pi_{t-1}(g^*))^2$$

where a_t is the atom the user provides feedback on and the expectation is taken over both the randomness of user's response and the randomness of structural QBC.

Proof. We first relate the variance of an atom to the probability of a mistake on that atom.

To this end, recall

$$D \;=\; \max_{a\in\mathcal{A}} \max_{g,g'\in\mathcal{G}} \|g(a) - g'(a)\|^2.$$

Moreover, since $\mathcal{G}$ and $\mathcal{A}$ are finite, we have the set of realizable values $S = \{g(a) : a \in \mathcal{A}, g \in \mathcal{G}\}$ is also finite. Let $d_o > 0$ denote the minimum squared distance between any two elements of S:

$$d_o \;=\; \min_{\substack{s,s'\in S:\\ s\neq s'}} \|s - s'\|^2.$$

For any atom $a \in \mathcal{A}$, we have

$$\frac{1}{D}\mathrm{var}(a;\pi_{t-1}) \;\leq\; \pi_{t-1}(\{g : g(a) \neq g^*(a)\}) \;\leq\; \frac{\mathrm{var}(a;\pi_{t-1})}{d_o\,\pi_{t-1}(g^*)}.$$

To see the left-hand inequality, we can work out

$$\begin{aligned}
\pi_{t-1}(\{g : g(a) \neq g^*(a)\}) &\geq \frac{1}{2}u(a;\pi_{t-1})\\
&= \frac{1}{2}\sum_{g,g'} \pi_{t-1}(g)\pi_{t-1}(g')\mathbf{1}[g(a) \neq g'(a)]\\
&\geq \frac{1}{2}\sum_{g,g'} \pi_{t-1}(g)\pi_{t-1}(g')\frac{\|g(a) - g'(a)\|^2}{D}\\
&= \frac{1}{D}\mathrm{var}(a;\pi_{t-1})
\end{aligned}$$

To see the right-hand inequality, notice

$$\begin{aligned}
\mathrm{var}(a;\pi_{t-1}) &= \frac{1}{2}\sum_{g,g'} \pi_{t-1}(g)\pi_{t-1}(g')\|g(a) - g'(a)\|^2\\
&\geq \pi_{t-1}(g^*)\sum_{g} \pi_{t-1}(g)\|g(a) - g^*(a)\|^2\\
&\geq \pi_{t-1}(g^*)\sum_{g} \pi_{t-1}(g)\, d_o\,\mathbf{1}[g(a) \neq g^*(a)]
\end{aligned}$$

$$\geq d_o\, \pi_{t-1}(g^*)\pi_{t-1}(\{g : g(a) \neq g^*(a)\})$$

Now suppose that structural QBC shows the user query q_t. Let $\epsilon > 0$ and $V_t = \text{var}(q_t; \pi_{t-1})$ be the random variable denoting the variance of q_t. For an atom $a \in A(q_t)$, we say that a has *low variance* if

$$\text{var}(a; \pi_{t-1}) \;<\; \epsilon V_t.$$

And we say that a has *high variance* otherwise. Then by a union bound, we have

$$\pi_{t-1}(\{g : g(a) \neq g^*(a) \text{ for some low variance } a \in A(q_t)\}) \;\leq\; \frac{\epsilon V_t |A(q_t)|}{d_o \pi_{t-1}(g^*)}.$$

On the other hand, there is *some* atom $a \in A(q_t)$ with at least average variance, so that

$$\pi_{t-1}(\{g : g(a) \neq g^*(a)\}) \;\geq\; \frac{\text{var}(a; \pi_{t-1})}{D} \;\geq\; \frac{V_t}{D}.$$

So with probability at least $V_t \left(\frac{1}{D} - \frac{\epsilon |A(q_t)|}{d_o \pi_{t-1}(g^*)} \right)$, a random draw from π_{t-1} gets all the low variance atoms correct and some high variance atom incorrect. Then conditioned on this event occurring, the user has probability at least p_o of correcting some high variance atom. So taking $\epsilon = \frac{d_o \pi_{t-1}(g^*)}{2DA_o}$ where $A_o = \max_q |A(q)|$, we have

$$\begin{aligned}\mathbb{E}\left[\text{var}(a_t; \pi_{t-1}) \,|\, \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\text{var}(a_t; \pi_{t-1}) | \text{var}(q_t; \pi_{t-1}) = V_t\right] \,|\, \mathcal{F}_{t-1}\right] \\ &\geq \mathbb{E}\left[\frac{V_t^2 p_o \epsilon}{2D} \,\middle|\, \mathcal{F}_{t-1} \right] \geq c_o \pi_{t-1}(g^*) \mathbb{E}[V_t^2 \,|\, \mathcal{F}_{t-1}] \geq c_o \pi_{t-1}(g^*) \mathbb{E}[V_t \,|\, \mathcal{F}_{t-1}]^2\end{aligned}$$

where $c_o = \frac{d_o p_o}{4D^2 A_o}$ is a positive constant. But by Lemma E.1, we know $\mathbb{E}[V_t \,|\, \mathcal{F}_{t-1}] \geq c\, \pi_{t-1}(g^*)(1 - \pi_{t-1}(g^*))$ for some constant $c > 0$. The lemma follows by substitution. □

Bibliography

[1] D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire de Probabilités XVII 1981/82*, pages 243–297. Springer, 1983.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[3] D. Angluin and L. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. In *Proceedings of the ninth annual ACM symposium on Theory of computing*, pages 30–41. ACM, 1977.

[4] S. Arora, R. Ge, and A. Moitra. Learning topic models: going beyond SVD. In *IEEE Symposium on Foundations of Computer Science*, 2012.

[5] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *33rd ACM Symposium on Theory of Computing*, pages 247–257, 2001.

[6] H. Ashtiani, S. Kushagra, and S. Ben-David. Clustering with same-cluster queries. In *Advances in Neural Information Processing Systems*, pages 3216–3224, 2016.

[7] P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 167–190, 2015.

[8] P. Awasthi, M.-F. Balcan, and K. Voevodski. Local algorithms for interactive clustering. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[9] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop. The hardness of approximation of Euclidean k-means. In *31st International Symposium on Computational Geometry*, volume 34, pages 754–767, 2015.

[10] P. Awasthi and R.B. Zadeh. Supervised clustering. In *Advances in Neural Information Processing Systems*, 2010.

[11] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[12] A. Baker. *Transcendental Number Theory.* Cambridge University Press, 1975.

[13] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[14] M.-F. Balcan and A. Blum. Clustering with interactive feedback. In *Algorithmic Learning Theory (volume 5254 of the series Lecture Notes in Computer Science)*, pages 316–328, 2008.

[15] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

[16] M-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Conference on Learning Theory*, pages 288–316, 2013.

[17] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, 2010.

[18] Y. Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.

[19] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, pages 192–236, 1974.

[20] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[21] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015.

[22] A. Bulatov and M. Grohe. The complexity of partition functions. *Theoretical Computer Science*, 348(2):148–186, 2005.

[23] L. M. Le Cam. On the speed of convergence of posterior distributions. Unpublished.

[24] E. Candes and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[25] R. Castro and R. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

[26] N. Cesa-Bianchi, C. Gentile, and F. Vitale. Learning unknown graphs. In *International Conference on Algorithmic Learning Theory*, pages 110–125, 2009.

[27] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear-threshold algorithms. In *Advances in Neural Information Processing Systems*, 2004.

[28] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[29] G. Dasarathy, R. Nowak, and X. Zhu. S2: An efficient graph based active learning algorithm with application to nonparametric classification. In *Proceedings of The 28th Conference on Learning Theory*, pages 503–522, 2015.

[30] S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, 2004.

[31] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, 2005.

[32] S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242, 2009.

[33] S. Dasgupta and D.J. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[34] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

[35] S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

[36] S. Dasgupta and M. Luby. Learning from partial correction. *Machine Learning Research*, 75:1–13, 2018.

[37] C. De Sa, C. Zhang, K. Olukotun, and C. Ré. Rapidly mixing Gibbs sampling for a class of factor graphs using hierarchy width. In *Advances in Neural Information Processing Systems*, pages 3079–3087, 2015.

[38] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[39] M. Dyer, L. A. Goldberg, C. Greenhill, and M. Jerrum. The relative complexity of approximate counting problems. *Algorithmica*, 38(3):471–500, 2004.

[40] J. Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The annals of applied probability*, pages 62–87, 1991.

[41] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168, 1997.

[42] A. Gelman, J.Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian data analysis*, volume 2. CRC Press, 2014.

[43] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

[44] R. Gilad-Bachrach, A. Navot, and N. Tishby. Query by committeee made real. In *Advances in Neural Information Processing Systems*, 2005.

[45] L. A. Goldberg and M. Jerrum. The complexity of ferromagnetic Ising with local fields. *Combinatorics, Probability and Computing*, 16(01):43–61, 2007.

[46] D. Golovin, A. Krause, and D. Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pages 766–774, 2010.

[47] A. Gotovos, H. Hassani, and A. Krause. Sampling from probabilistic submodular models. In *Advances in Neural Information Processing Systems*, pages 1936–1944, 2015.

[48] A. Guillory and J. Bilmes. Average-case active learning with costs. In *International Conference on Algorithmic Learning Theory*, pages 141–155, 2009.

[49] V. Guruswami and A. Vardy. Maximum-likelihood decoding of Reed-Solomon codes is NP-hard. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 470–478, 2005.

[50] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished, 1971.

[51] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 25th International Conference on Machine Learning*, 2007.

[52] M. Hardt and E. Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing*, pages 753–760, 2015.

[53] N. J. Higham. *Accuracy and stability of numerical algorithms.* SIAM, 2002.

[54] G. Hinton, S. Osindero, and Y-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[55] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[56] D. Hsu and S.M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Fourth Innovations in Theoretical Computer Science*, 2013.

[57] Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

[58] E. Ising. Report on the theory of ferromagnetism. *Zeitschrift fur physik*, 31:253–258, 1925.

[59] J. Jawitz. Moments of truncated continuous univariate distributions. *Advances in water resources*, 27(3):269–281, 2004.

[60] M. Jerrum. *Counting, sampling and integrating: algorithms and complexity.* Springer Science & Business Media, 2003.

[61] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

[62] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[63] A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

[64] S. Kpotufe, R. Urner, and S. Ben-David. Hierarchical label queries with data-dependent partitions. In *Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

[65] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[66] D. Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, 2008.

[67] M. Lichman. UCI machine learning repository, 2013.

[68] X. Liu and J. Domke. Projecting Markov random field parameters for fast mixing. In *Advances in Neural Information Processing Systems*, pages 1377–1385, 2014.

[69] P. M. Long and R. A. Servedio. Restricted Boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27th International Conference on Machine Learning*, pages 703–710, 2010.

[70] L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30:307–358, 2007.

[71] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13–21, 2012.

[72] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 93–102, 2010.

[73] K. Murphy. *Machine learning: a probabilistic perspective.* MIT Press, 2012.

[74] V. Nair and G. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

[75] K. W. Ng, G.-L. Tian, and M.-L. Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications.* Wiley, 2011.

[76] R. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

[77] R. B. Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the Cambridge philosophical society*, volume 48, pages 106–109, 1952.

[78] L. Rüschendorf and S. T. Rachev. A characterization of random variables with minimum l2-distance. *Journal of Multivariate Analysis*, 32(1):48–54, 1990.

[79] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. In *International conference on artificial intelligence and statistics*, pages 448–455, 2009.

[80] B. Settles. *Active learning.* Morgan Claypool, 2012.

[81] H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 287–294, 1992.

[82] A. Sinclair. *Randomised Algorithms for Counting and Generating Combinatorial Structures.* PhD thesis, University of Edinburgh, 1988.

[83] D. Štefankovič, Santosh S. Vempala, and E. Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM (JACM)*, 56(3):18, 2009.

[84] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

[85] S. Vikram and S. Dasgupta. Interactive Bayesian hierarchical clustering. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[86] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[87] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[88] N-Y. Wang and L. Wu. Convergence rate and concentration inequalities for Gibbs sampling in high dimension. *Bernoulli*, 20(4):1698–1716, 2014.

[89] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.

[90] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML Workshop on the Continuum from Labeled to Unlabeled Data*, 2003.