Title

Analysis of Dictionary Learning and Random Forest under Data-inspired Models

Author

Wang, Yu

Publication Date

2020

Peer reviewed|Thesis/dissertation

Analysis of Dictionary Learning and Random Forest under Data-inspired Models

by

Yu Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Chair
Professor Adityanand Guntuboyina
Professor Thomas Courtade

Fall 2020

Analysis of Dictionary Learning and Random Forest under Data-inspired Models

Abstract

Analysis of Dictionary Learning and Random Forest under Data-inspired Models

by

Yu Wang

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Chair

Many algorithms in Machine Learning have demonstrated powerful empirical performance in terms of prediction accuracy and ability to extract knowledge from data. Inspired by the empirical success, researchers study the behavior of those algorithms under theoretical models. Two challenges must be addressed when one tries to understand those algorithms from a theoretical perspective. First, a sound theoretical model should be considered. A good model should reflect key properties of the real data. If the models do not capture important aspects of the real data, the observations and/or conclusions made on those theoretical models may not be relevant to practice. Thus, theoretical models that reflect certain properties of real data are required so that the insights obtained from those models are convincing to practitioners. Second, many of these algorithms are hard to analyze via traditional techniques. Therefore, novel techniques are required to analyze those algorithms under new theoretical models. In this thesis, we analyze two problems under novel real-data inspired theoretical models: ℓ_1-minimization for dictionary learning and seeking important features and feature interactions from Random Forest (RF). Both analyses give unique insights into the problem by studying a novel data generative model. For dictionary learning, we propose two novel theoretical models: exact sparse model and Bernoulli-type models. Unlike most previous analyses that assume the data is generated from Gaussian distributions with sparsity constraints, these new models can capture non-Gaussian data distributions and allow us to analyze the algorithms under novel data properties such as non-negativity and heavy-tail. We show that ℓ_1-minimization model in Dictionary Learning does not satisfy the classic global identifiability condition under the new model. However, the reference dictionary still enjoys some global property across all the feasible dictionaries. Our theoretical analysis leads to a novel algorithm called Dictionary Learning Block Coordinate Descent (DL-BCD). For RF, we start off with analyzing the feature importance bias for noisy features when using Mean Decrease Impurity (MDI). Then, we study the feature interaction recovery problem and analyzed the data-inspired Local-Spiky Sparse (LSS) model without Lipschitz assumptions that are often present in the previous literature. We show that the depth-weighted-prevalence of a true feature interaction in the

decision paths of trees does not depend on the model coefficients but only on the size of the interaction. The theoretical analysis leads to a novel feature ranking method called LSSrank. We examine the performance of LSSrank on simulated data and it has high probability to rank true interactions at the top under the LSS model.

To my family

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I owe a great debt to my Ph.D. advisor, Bin Yu. In the course of the past five years, Bin not only guided me with her sharp vision in many research projects, but also taught me her predictivity-computability-stability (PCS) principles that would benefit me in future work and research. Through her STAT 215A class, Bin led me into the wonderland of practical statistics and taught me many useful knowledge on how to solve real data problems. Bin is also a wonderful role model. Bin gave me many sincere and insightful advice on how to improve myself in research and personal development. Bin invites all of her students to her home every Thanksgiving and it has been one of my greatest experiences at Berkeley. I feel fortunate to be part of Bin's research group Yu-group, which Bin spends a tremendous amount of time nurturing every member. Without Bin's guidance, I would not be able to have such a rewarding and memorable experience at Berkeley.

I have learned the most through interacting with Yu group members and alumni. I appreciate Xiangyu Chang and Siqi Wu who first introduced me to the group and brought me on board. It has been a lot of fun working with people in our group. I received a lot of help from group members when we conducted research together including Siqi Wu, Yan Shuo Tan, Xiao Li, Merle Behr, Tiffany Tang, Chandan Singh, and Reza Abbassi. I would also like to thank all the kind help and fun conversations I had from other group members: Christine Kuang, Wooseok Ha, Raaz Dwivedi, Simon Walter, Yuansi Chen, Nick Altieri, Rebecca Barter, Abhineet Agarwal, Robert Netzorg, Tiffany Tang, Soeren Kunzel, Hanzhong Liu, Sumanta Basu, Shamindra Shrotriya, and Briton Park.

I appreciate all the knowledge I learned from awesome faculty at Berkeley through courses and projects. Peng Ding and Sam Pimentel introduced me to causal inference. Phillip Stark taught me a lot about statistics used in real life such as voting and about how to critically read papers. I also learned a lot of statistics and machine learning from Adityanand Guntuboyina, Will Fithian, David Aldous, Martin Wainwright, and Mike Jordan.

I appreciate my valuable experience at two summer internships at Microsoft and Two Sigma. I learned practical knowledge about causal inference and A/B testing at ExP team at Microsoft. I would like to thank my manager Benjamin Arai, my mentor Somit Gupta, my collegues Jiannan Lu, Ali Mahmoudzadeh, and Sophia Liu. I also learned about how to analyze financial data at Two Sigma. For that, I would like to thank my manager Bo Jiang and colleague Jeffrey Ning.

I am grateful of the love and support I got from my wife Shiying He during my PhD. I appreciate all the joys and tears we went through together that made this challenging PhD career so much more lovely and enjoyable. I would like to thank my parents for all the love, support, and understanding I received in the past 28 years. I would like to thank my friends outside the group: Hengke Han, Yang An, Rui Gao, Hamza Hussain, Sam Jafarian, Ryan Giordano, Samuel Lin, Tom Chou, Jason Wu, Yuting Wei, Yuting Ye, Yumeng Zhang, Lihua Lei, Jianbo Chen, Zhiyi You, Junyu Cao, Wilson Cai, Zsolt Bartha, and Gerry Zhang.

I am grateful of so many people who helped me to accomplish this journey. I would like to thank my committee members: Bin Yu, Adityanand Guntuboyina, Thomas Courtade, and

Sam Pimentel. I appreciate all the help I got from La Shana Polaris, Mary Melinn, and other staff at our department. Before coming to Berkeley, many professors and teachers kindly spent so much efforts nurturing me and directed me to this long yet rewarding path: Jinhua Gong, Shijun Zhang, Chunhua Qin, Lei Jin, Tian Li, Wotao Yin, Jinshan Zeng, Zongben Xu, and Tom Chou.

Chapter 1

Introduction

Machine Learning algorithms have been successfully applied to many real-world problems such as image processing and genomics [61, 12, 82, 10]. Though the machine learning models have higher predictive power than the traditional methods, they are usually viewed as black boxes and it is difficult to understand how and why they achieve high predicative power. To go beyond their plain empirical success and pursue a deeper understanding of when a machine learning model will work or break, it is important for researchers to investigate an algorithm's behavior under some theoretical model. However, a theoretical model is a double-edged sword. A carefully chosen theoretical model can capture the key properties of a real data set and sheds insight on the behavior of an algorithm. Metaphorically, a good model serves as a microscope that zoom into the key aspects of the algorithm and extracts hidden information. On the other hand, a bad model does not capture properties of real data. As a result, lessons learned from studying those models do not readily provide insights into the algorithm performance on a real data set.

Studying an algorithm under a new theoretical model is often not easy. First of all, many machine learning algorithms are quite complex and considered a black box. From optimization point of view, they are usually non-convex, which means their behavior can be affected by many seemingly mundane things such as the initialization. Also, techniques developed for classic theoretical models are often not able to be carried over to analyze a new theoretical model. There is also less literature one can rely on. As a result, novel techniques are required to analyze those algorithms under new theoretical models. Due to limitations of the available resources for analyzing new theoretical models, there is usually a trade-off between utility, i.e., how well a model approximates reality, and feasibility, i.e., whether it is feasible to analyze the algorithm under such a model.

In this thesis, we analyze two problems: ℓ_1-minimization for dictionary learning and seeking important features and feature interactions via RF. Both cases show the importance of studying an ML algorithm under a data inspired model, allowing one to draw many useful insights.

1.1 Local and Global identifiability of dictionary learning

In the first part of thesis (Chapter 2–6) we study the problem of globally recovering a dictionary from a set of signals via ℓ_1-minimization. We assume that the signals are generated as i.i.d. random linear combinations of the K atoms from a complete reference dictionary $\boldsymbol{D}^* \in \mathbb{R}^{p\times p}$, where the linear combination coefficients are from either a Bernoulli type model or exact sparse model. First, we obtain a necessary and sufficient norm condition for the reference dictionary $\boldsymbol{D}^*$ to be a sharp local minimum of the expected ℓ_1 objective function. Our result substantially extends that of [100] and allows the combination coefficient to be non-negative. Second, we obtain an explicit bound on the region within which the objective value of the reference dictionary is minimal. Third, we show that the reference dictionary is the unique sharp local minimum, thus establishing the first known global property of ℓ_1-minimization dictionary learning. Motivated by the theoretical results, we introduce a perturbation based test to determine whether a dictionary is a sharp local minimum of the objective function. In addition, we also propose a new dictionary learning algorithm based on Block Coordinate Descent, called DL-BCD, which is guaranteed to decrease the obective function monotonically. Simulation studies show that DL-BCD has competitive performance in terms of recovery rate compared to other state-of-the-art dictionary learning algorithms when the reference dictionary is generated from random Gaussian matrices.

The chapters are organized as follows. We give a high-level review of the related works and summarize our theoretical contributions in Chapter 2. In Chapter 3, we introduce necessary notations and propose two novel theoretical models: exact sparse model and Bernoulli-type models. These new models can capture non-Gaussian data distributions and allow us to analyze the algorithms under novel data properties such as non-negativity and heavy-tail. In Chapter 4, we present main theorems and discuss their implications. In Chapter 5, we propose the sharpness test and the block coordinate descent algorithm for dictionary learning (DL-BCD) and evaluate their numerical performance. We conclude our results and discuss possible extensions in Chapter 6.

1.2 Feature importance and feature interaction discovery via RF

Random Forest (RF) is at the cutting edge of supervised machine learning methods especially for genomics problems, and for biological interaction discovery as by stabilized RF or iterative random forest (iRF)[10]. We introduce the problem setup of feature importance and feature interaction of RF in Chapter 7. We study how to compute RF feature importance in Chapter 8 and how to use RF to discover high-order feature interactions in Chapter 9 – Chapter 11.

RF feature importance

To understand how tree ensembles make predictions, people routinely turn to feature importance measures calculated from tree ensembles. It has long been known that Mean Decrease Impurity (MDI), one of the most widely used measures of feature importance, incorrectly assigns high importance to noisy features, leading to systematic bias in feature selection. We address the bias of MDI from both theoretical and methodological perspectives. Based on the original definition of MDI by Breiman et al.[13] for a single tree, we derive a tight non-asymptotic bound on the expected bias of MDI importance of noisy features, showing that deep trees have higher (expected) feature selection bias than shallow ones. However, it is not clear how to reduce the bias of MDI using its existing analytical expression. We derive a new analytical expression for MDI, and based on this new expression, we are able to propose a debiased MDI feature importance measure using out-of-bag samples, called MDI-oob. For both the simulated data and a genomic ChIP dataset, MDI-oob achieves state-of-the-art performance in feature selection from RF for both deep and shallow trees.

Feature interaction discovery

There is no theoretical results on how to use tree-based methods to find high-order feature interactions in the existing literature. We propose a new model, the Locally Spiky Sparse (LSS) regression model, which is biologically inspired without Lipschitz assumptions. The regression function in LSS is a linear combination of a set of piece-wise constant, discontinuous Boolean local interaction functions. We show that (with high probability) the depth-weighted prevalence (DWP) of interactions among decision paths of RF is universal (i.e., independent of any model coefficients) and only depends on the size of the interactions. Our theoretical analysis reveals for the first time that the feature subsampling strategy used in RF, i.e., splitting each node of a tree with a subset of mtry candidate features, is key to obtain exact interaction recovery. Inspired by this theoretical result, we propose a novel method, namely LSSrank, to rank high-order interactions based on DWP. We conduct a series of simulations on synthetic data and biologically-inspired data, and find that LSSrank gives correct ranking results with high probability under the LSS model.

Part I

Local and global identifiability in dictionary learning

Chapter 2

Dictionary learning and ℓ_1-minimization

Dictionary learning is a class of unsupervised learning algorithms that learn a data-driven representation from signals such as images, speech, and video. It has been widely used in many applications ranging from image imputation to texture synthesis [77, 59, 71]. Compared to pre-defined dictionaries, data-driven dictionaries can extract meaningful and interpretable patterns from scientific data [65, 66] and exhibit enhanced performance in blind source separation, image denoising and matrix completion. See, e.g., [104, 43, 48, 25, 3, 61, 73] and the references therein. Dictionary learning is also closely related to Non-negative Matrix Factorization (NMF) [46] which has broad applications in biology [14, 101]. Despite many successful applications, dictionary learning formulations and algorithms are generally hard to analyze due to their non-convex nature. With different initial inputs, a dictionary learning algorithm typically outputs different dictionaries as a result of this non-convexity. For those who use the dictionary as a basis for downstream analyses, the choice of the dictionary may significantly impact the final conclusions. Therefore, it is natural to ask the following questions: Can dictionary learning algorithms recover the "ground-truth" dictionary if there is one? Among the many outputs from a dictionary learning algorithm, which one should be selected for further analysis? In this chapter, we give a high level overview of the existing literature and introduce our methodology and contributions.

Literature review

To answer the above questions, we need to understand the theoretical properties of dictionary learning under generative models. In a number of recent works, the signals are generated as linear combinations of the columns of a reference dictionary [32, 30, 31]. Specifically, denoting by $\boldsymbol{D}^* \in \mathbb{R}^{d\times K}$ the reference dictionary and $\boldsymbol{x}^{(i)} \in \mathbb{R}^d, i = 1, \ldots, n$ the signal vectors, we have:

$$\boldsymbol{x}^{(i)} \approx \boldsymbol{D}^* \boldsymbol{\alpha}^{(i)}, \tag{2.1}$$

where $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^K$ denotes the sparse coefficient vector. If $K = d$ and $\boldsymbol{D}^*$ is full rank, the dictionary is called *complete*. If the matrix has more columns than rows, i.e., $K > d$, the

dictionary is *overcomplete*. Under the model (2.1), for any reasonable dictionary learning objective function, the reference dictionary $\boldsymbol{D}^*$ ought to be equal or close to a local minimum. This wellposedness requirement, also known as *local identifiability* of dictionary learning, turns out to be nontrivial. For a complete dictionary and noiseless signals, [32] studies the following ℓ_1-minimization formulation:

$$\text{minimize}_{\boldsymbol{D},\{\boldsymbol{\beta}^{(i)}\}_{i=1}^n} \sum_{i=1}^{n} \|\boldsymbol{\beta}^{(i)}\|_1 \tag{2.2}$$

$$\text{subject to } \|\boldsymbol{D}_j\|_2 \leq 1, j = 1, \ldots, K,$$

$$\boldsymbol{x}^{(i)} = \boldsymbol{D}\boldsymbol{\beta}^{(i)},\ i = 1, \ldots, n; \tag{2.3}$$

where $\boldsymbol{D}$ is a complete dictionary and $\boldsymbol{D}_j$ is its j-th column. [32] proved a sufficient condition for local identifiability under the Bernoulli-Gaussian model. A more refined analysis by [100] gave a sufficient and almost necessary condition. The sufficient local identifiability condition in [32] was extended to the over-complete case [30] and the noisy case [31].

As most of dictionary learning formulations are nonconvex, local identifiability alone does not guarantee that the output dictionary is the reference dictionary — the initial dictionary must also be quite close to the reference dictionary. There are only limited results on how to choose an appropriate initialization. For example, [7] showed that their initialization algorithm guarantees that the output dictionary is within a small neighborhood of the reference dictionary when certain μ-incoherence condition is met. In practice, initialization is usually done by using a random matrix or randomly selecting a sample of signals [60]. These algorithms are typically run for multiple times and the dictionary that achieves the smallest objective value is selected.

The difficulty of initialization is a major challenge of establishing the recovery guarantee that under some generative models, the output dictionary of an algorithm is indeed the reference dictionary. This motivates the study of *global identifiability*. There are two versions of global identifiability. For the first version, we say that the reference dictionary $\boldsymbol{D}^*$ is globally identifiable with respect to an objective function $L(\cdot)$ if $\boldsymbol{D}^*$ is a global minimum of L. The second and stricter version, requires all local minima of L are the same as $\boldsymbol{D}^*$ up to column sign changes and permutation. If the second version of global identifiability holds, all local minima are global minima. Thus any algorithm capable of converging to a local minimum will also recover the reference dictionary. For some matrix decomposition tasks such as low rank PCA [85] and matrix completion [29], despite the fact that the objective function is non-convex, the stricter version of global identifiability holds under certain conditions. For dictionary learning, several papers proposed new algorithms with theoretical recovery guarantees that ensure the output is close or equal to the reference dictionary. For the complete and noiseless case, [84] proposed a linear programming based algorithm that provably recovers the reference dictionary when the coefficient vectors are generated from a Bernoulli Gaussian model and contain at most $O(\sqrt{K})$ nonzero elements. [90, 91] improved the sparsity tolerance to $O(K)$ using a Riemannian trust region method.

For over-complete dictionaries, [5] proposed an algorithm which performs an overlapping clustering followed by an averaging algorithm or a K-SVD type algorithm. Additionally, there is another line of research that focuses on the analysis of alternating minimization algorithms, including [1, 2, 6, 7]. [9] proposed an algorithm based on sum-of-square semi-definite programming hierarchy and proved its desirable theoretical performance with relaxed assumptions on coefficient sparsity under a series of moment assumptions.

Our contributions

Despite numerous studies of global recovery in dictionary learning, our result is the first global identifiability result for the ℓ_1-minimization problem. As we illustrate in Chapter 4, the reference dictionary may not be the global minimum even for a simple data generation model. This motivates us to consider a different condition to distinguish the reference dictionary from other local minima. We show that the reference dictionary is the unique "sharp" local minimum (see Definition 3.2.1) of the ℓ_1 objective function when certain conditions are met – in other words, there are no other sharp local minima than the reference dictionary.

Based on this new characterization and the observation that a sharp local minimum is more resilient to small perturbations, we propose a method to empirically test the sharpness of the objective function at a given dictionary. Furthermore, we also design a new algorithm to solve the ℓ_1-minimization problem using Block Coordinate Descent (DL-BCD) and the re-weighting scheme inspired by [15]. Our simulations demonstrate that the proposed method compares favorably with other state-of-the-art algorithms in terms of recovery rate if the reference dictionary is generated from random Gaussian matrices.

Our work differs from other recent studies in two main aspects. First, instead of proposing new dictionary learning formulations, we study the global property of the existing ℓ_1-minimization problem that is often considered difficult in previous studies [61, 100]. While there are many dictionary learning algorithms that do not rely on the ℓ_1-type penalty, formulations with ℓ_1 penalties remain as the most frequently used method in many applications due to their good practical performance and the availability of efficient algorithms [61, 59]. The theoretical understanding of ℓ_1-minimization is therefore of interest to a wider audience than other dictionary learning methods. Second, our data generation models are novel and cover several important cases not studied by prior works, e.g., non-negative linear coefficients. Even though there is a line of research that focuses on non-negative dictionary learning in the literature [4, 37, 6], the reference dictionary and the corresponding coefficients therein are both non-negative. In comparison, we allow the dictionary to have arbitrary values but only constrain the reference coefficients to be non-negative. This non-negative coefficient case is difficult to analyze and does not satisfy the recovery conditions in previous studies, for instance [9, 90, 91].

Chapter 3

Exact sparse models and Bernoulli-type models

3.1 Preliminaries

For a vector $\boldsymbol{w} \in \mathbb{R}^m$, denote its j-th element by w_j. For an arbitrary matrix $A \in \mathbb{R}^{m\times n}$, let $A[k,], A_j, A_{k,j}$ denote its k-th row, j-th column, and the (k,j)-th element respectively. Denote by $A[k,-j] \in \mathbb{R}^{n-1}$ the k-th row of A without its j-th entry. Let $\mathbb{I} \in \mathbb{R}^{K\times K}$ denote the identity matrix of size K and for $k \in \{1,\ldots,K\}$, $\mathbb{I}_k$ denotes $\mathbb{I}$'s k-th column, whose k-th entry is one and zero elsewhere. $\mathbf{1} \in \mathbb{R}^{K\times 1}$ denotes a column vector whose elements are all ones. For a positive semi-definite square matrix $X \in \mathbb{R}^{K\times K}$, $X^{1/2}$ denotes its positive semi-definite square root. We use $\|\cdot\|$ to denote vector norms and $|||\cdot|||$ to denote matrix (semi-)norms. In particular, $|||\cdot|||_F$ denotes the Frobenius norm, whereas $|||\cdot|||_2$ denotes the spectral norm. For any two real functions $w(t), q(t) : \mathbb{R} \to \mathbb{R}$, we denote $w(t) = \Theta(q(t))$ if there exist constants $c_1, c_2 > 0$ such that for any $t \in \mathbb{R}$, $c_1 < \frac{w(t)}{q(t)} < c_2$. If $q(t) > 0$ and $\lim_{q(t)\to 0} \frac{w(t)}{q(t)} = 0$, then we write $w(t) = o(q(t))$. Define the indicator and the sign functions as

$$\mathbf{1}(x=0) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}, \ \operatorname{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}.$$

The ℓ_1 dictionary learning objective

In dictionary learning, a dictionary is represented by a matrix $\boldsymbol{D} \in \mathbb{R}^{d\times K}$. We call a column of the dictionary matrix an atom of the dictionary. In this thesis, we consider complete dictionaries, that is, the dictionary matrix is square ($K = d$) and invertible. Note that for the noiseless case, an undercomplete dictionary ($K < d$) can always be reduced to a complete dictionary by removing certain rows. A complete or undercomplete dictionary matrix is typically used in applications such as Independent Component Analysis [21] and Non-negative Matrix Factorization [46, 14, 101].

For a complete dictionary $\boldsymbol{D}$, define L as the ℓ_1 objective function:

$$L(\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\beta}^{(i)}\|_1. \quad \text{where } \boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)} \ \forall\, i \in 1, \ldots, n. \tag{3.1}$$

The ℓ_1-minimization formulation (2.2) is equivalent to the following optimization problem [100]:

$$\text{minimize}_{\boldsymbol{D}\in\mathbb{B}(\mathbb{R}^K)} L(\boldsymbol{D}), \tag{3.2}$$

where $\mathbb{B}(\mathbb{R}^K)$ is the set of all feasible dictionaries:

$$\mathbb{B}(\mathbb{R}^K) \triangleq \left\{ \boldsymbol{D} \in \mathbb{R}^{K\times K} \middle| \|\boldsymbol{D}_1\|_2 = \ldots = \|\boldsymbol{D}_K\|_2 = 1, \ \text{rank}(\boldsymbol{D}) = K \right\}. \tag{3.3}$$

3.2 Generative models

Let $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ be the reference dictionary of interest. We assume that the signal vector $\boldsymbol{x} \in \mathbb{R}^K$ is generated from a linear model without noise: $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^K$ is a random reference coefficient vector. Below, we will introduce two classes of generative models for $\boldsymbol{\alpha}$: Bernoulli-type models and exact sparse models.

- *Bernoulli-type model* $\mathcal{B}(p_1, \ldots, p_K; f)$. Let $\boldsymbol{z} \in \mathbb{R}^K$ be a random vector whose probability density function exists and is denoted by f. Let $\boldsymbol{\xi} \in \{0,1\}^K$ be a random boolean vector. The coordinates of $\boldsymbol{\xi}$ are independent and ξ_j is a Bernoulli random variable with success probability $P(\xi_j = 1) = p_j \in (0,1)$. Define $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\alpha_j = \xi_j z_j$ for all j. We say that $\boldsymbol{\alpha}$ is generated from the Bernoulli-type model $\mathcal{B}(p_1, \ldots, p_K; f)$.

- *Exact sparse model* $\mathcal{S}(s; f)$. Let $\boldsymbol{z} \in \mathbb{R}^K$ be a random vector whose probability density function exists and is denoted by f. Let $\boldsymbol{S}$ be a size-s subset uniformly drawn from all size-s subsets of $1, \ldots, K$. Let $\boldsymbol{\xi} \in \{0,1\}^K$ be a random variable such that $\xi_j = 1$ if $j \in \boldsymbol{S}$ otherwise 0. Define $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\alpha_j = \xi_j z_j$ for all j. We say that $\boldsymbol{\alpha}$ is generated from the exact sparse model $\mathcal{S}(s; f)$.

These two classes can be viewed as natural extensions of Bernoulli Gaussian models and sparse Gaussian models, which have been extensively studied in dictionary learning [32, 100, 80, 81]. Denote by $\mathcal{N}(0, \mathbb{I}_{k\times k})$ the k-dimensional standard Gaussian distribution:

- *Bernoulli Gaussian model.* If $\boldsymbol{\alpha}$ is generated from the Bernoulli-type model with parameters $p_j = p$ $(p > 0)$ for all j and f is the density of $\mathcal{N}(0, \mathbb{I}_{k\times k})$, we say that $\boldsymbol{\alpha}$ follows a Bernoulli Gaussian model with parameter p, or $BG(p)$.

- *Sparse Gaussian model.* If $\boldsymbol{\alpha}$ is generated from the exact sparse model with sparsity parameter s and f is the density of $\mathcal{N}(0, \mathbb{I}_{k\times k})$, we say that $\boldsymbol{\alpha}$ follows the sparse Gaussian model with parameter s, or $SG(s)$.

Remarks: The advantage of using sparse Gaussian and Bernoulli Gaussian distributions is that they are simple and yet capable of capturing the most important characteristic of the reference coefficients: sparsity. By using sparse Gaussian and Bernoulli Gaussian distributions, Wu and Yu[100] obtains a sufficient and almost necessary condition for local identifiability. Take the sparse Gaussian distribution as an example: let the maximal collinearity μ of the reference dictionary $\boldsymbol{D}^*$ be $\mu = \max_{i \neq j} \left| \boldsymbol{D}_i^{*T} \boldsymbol{D}_j^* \right|$ and s be the sparsity of the reference coefficient vector in the sparse Gaussian model. Wu and Yu[100] show that local identifiability holds when $\mu < \frac{K-s}{\sqrt{s}(K-1)}$. From the formula, we can see a trade-off between the maximal collinearity μ and the sparsity of the coefficient vector s. If the coefficient is very sparse, i.e., $s \ll K$, local identifiability holds for a wide range of μ. Otherwise, local identifiability holds for a narrower range of μ. While sparse/Bernoulli Gaussian models can be used to illustrate this trade-off, they are rather restrictive for real data. Several papers [84, 6, 5, 7, 31] studied more general models such as sub-Gaussian models.

Other important examples include models with z drawn from the Laplacian distribution or a non-negative distribution. In particular, the non-negativity of the coefficients violates the popular assumption $\mathbb{E}\boldsymbol{\alpha}_j = 0$ [32, 31].

- *Sparse Laplacian model.* If $\boldsymbol{\alpha}$ is generated from the exact sparse model with sparsity parameter s and density $f(z) = \frac{1}{2^K} \exp(-\|z\|_1)$, we say that $\boldsymbol{\alpha}$ follows the sparse Laplacian model with parameter s, or $SL(s)$.

- *Non-negative Sparse Gaussian model.* A random vector $\boldsymbol{\alpha}$ is said to be drawn from a non-negative sparse Gaussian model with parameter s, denoted by $|SG(s)|$, if for $j = 1, \ldots, K$, $\boldsymbol{\alpha}_j = |\boldsymbol{\alpha}'_j|$ where $\boldsymbol{\alpha}' \sim SG(s)$.

Identifiability of the reference dictionary

In this subsection, we introduce commonly used terminology in dictionary learning with respect to the identifiability of the reference dictionary.

- *Sign-permutation ambiguity.* In most dictionary learning formulations, the order of the dictionary atoms as well as their signs do not matter. Let $P \in \mathbb{R}^{K \times K}$ be a permutation matrix and $\Lambda \in \mathbb{R}^{K \times K}$ a diagonal matrix with ± 1 diagonal entries. The matrix $\boldsymbol{D}' = \boldsymbol{D}P\Lambda$ and $\boldsymbol{D}$ essentially represent the same dictionary but $\boldsymbol{D}' \neq \boldsymbol{D}$ element-wise.

- *Local identifiability.* The reference dictionary $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ is *locally identifiable* with respect to L if $\boldsymbol{D}^*$ is a local minimum of L. Local identifiability is a minimal requirement for recovering the reference dictionary. It has been extensively studied under a variety of dictionary learning formulations [32, 30, 31, 100, 2, 81].

- *Global identifiability.* The reference dictionary $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ is *globally identifiable* with respect to L if $\boldsymbol{D}^*$ is a global minimum of L.

Clearly, whether local or global identifiability holds depends on the objective function and the signal generation model. If the objective function is ℓ_0, i.e., $\frac{1}{n}\sum_i \|\boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}\|_0$, and the linear coefficients are generated from the Bernoulli Gaussian model, the reference dictionary is globally (and hence locally) identifiable (see Theorem 3 in [84]). However, for the ℓ_1 objective considered here, global identifiability might not hold. In Chapter 4, we give an example where the reference dictionary is only a local minimum but not a global minimum.

We consider a variant of global identifiability: instead of the global minimum, we require the reference dictionary $\boldsymbol{D}^*$ to be the *unique sharp local minimum* of the dictionary learning objective function. In other words, no dictionary other than $\boldsymbol{D}^*$ is a sharp local minimum. Other dictionaries can still be local minima but cannot be *sharp* at the same time. This property allows us to globally distinguish the reference dictionary from other spurious local minima and can be used as a criterion to select the best dictionaries from a set of algorithm outputs. Sharp local minimum, as per Definition 3.2.1, is a common concept in the field of optimization [23, 72]. However, to the best of our knowledge, we are the first to connect dictionary learning theory with sharp local minimum and use it to distinguish the reference dictionary from other spurious local minima.

Definition 3.2.1 (Sharp local minimum)**.** *Let $L : \mathbb{B}(\mathbb{R}^K) \to \mathbb{R}$ be a dictionary learning objective function. A dictionary $\boldsymbol{D}^0 \in \mathbb{B}(\mathbb{R}^K)$ is a sharp local minimum of $L(\cdot)$ with sharpness ϵ [72] if there exists $\delta > 0$ such that for any $\boldsymbol{D} \in \{\boldsymbol{D} : |||\boldsymbol{D} - \boldsymbol{D}^0|||_F < \delta\}$:*

$$L(\boldsymbol{D}) - L(\boldsymbol{D}^0) \geq \epsilon|||\boldsymbol{D} - \boldsymbol{D}^0|||_F + o(|||\boldsymbol{D}^0 - \boldsymbol{D}|||_F).$$

Remarks: The definition here can be viewed as a matrix analog of the sharp minimum in the one dimensional case. For a function $f : \mathbb{R} \to \mathbb{R}$, v^0 is a sharp local minimum of f if $f(v) - f(v^0) \geq \epsilon|v - v^0| + o(|v - v^0|)$. Note that the definition of sharp local minimum is different from the definition of strict local minimum, which means there are no other local minima in its neighborhood. A sharp local minimum is always a strict local minimum but not vice versa. For example, consider ℓ_q functions $|x|^q$ for $q > 0$. When $q \leq 1$, $x = 0$ is a strict local minimum as well as a sharp local minimum of ℓ_q. When $q > 1$, $x = 0$ is still a strict local minimum but not a sharp local minimum. This definition is also different from the sharp local minimum concepts that are commonly used in the study of artificial neural networks and stochastic gradient descent [34].

Technical assumptions

In this subsection, we introduce two important technical assumptions that will be used in our theoretical analysis. All the models introduced in Section 3.2 satisfy these assumptions. Their relationship is depicted in Fig. 3.1.

We need additional notations before introducing the assumptions. For any $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)$, define $M(\boldsymbol{D}) = \boldsymbol{D}^T\boldsymbol{D}$ as the collinearity matrix of $\boldsymbol{D}$. For example, if the dictionary is an

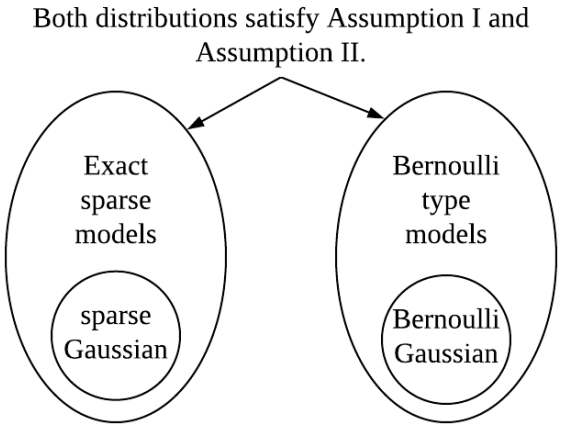


Figure 3.1: Both exact sparse models and Bernoulli-type models satisfy Assumption I and II. Sparse Gaussian distribution is a special case of exact sparse models, while Bernoulli Gaussian distribution is a special case of Bernoulli-type models.

orthogonal matrix, $M(\boldsymbol{D}) = \mathbb{I}$ is the identity matrix. If all the atoms in the dictionary are collinear with constant $\mu > 0$, then $M(\boldsymbol{D}) = \mu \mathbf{1}\mathbf{1}^T + (1-\mu)\mathbb{I}$ is a matrix whose off-diagonal elements are all μ's. When the context is clear, we use M instead of $M(\boldsymbol{D})$ for notation ease. Denote by $M^* \equiv M^*(\boldsymbol{D}^*)$ the collinearity matrix for the reference dictionary $\boldsymbol{D}^*$. Also, define the matrix $B(\boldsymbol{\alpha}, M) \in \mathbb{R}^{K\times K}$ as

$$(B(\boldsymbol{\alpha}, M))_{k,j} \triangleq \mathbb{E}\boldsymbol{\alpha}_j \text{sign}(\boldsymbol{\alpha}_k) - M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| \quad \text{for } k, j = 1, \ldots, K.$$

Here the expectation is with respect to the random coefficient vector $\boldsymbol{\alpha}$. By the definition of $B(\boldsymbol{\alpha}, M^*)$, the quantity is the difference between two matrices

$$B(\boldsymbol{\alpha}, M^*) = B_1(\boldsymbol{\alpha}) - B_2(\boldsymbol{\alpha}, M^*),$$

where $(B_1(\boldsymbol{\alpha}))_{k,j} = \mathbb{E}\boldsymbol{\alpha}_j \text{sign}(\boldsymbol{\alpha}_k)$ and $B_2(\boldsymbol{\alpha}, M^*)_{k,j} = M^*_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j|$. Roughly speaking, the first matrix measures the "correlation" between different coordinates of the coefficients while the second matrix measures the collinearity of the atoms in the reference dictionary. For instance, when the coordinates of $\boldsymbol{\alpha}$ are independent and mean zero, $B_1(\boldsymbol{\alpha}) = 0$. When all atoms in the dictionary are orthogonal, i.e., $M^* = \mathbb{I}$, $B_2(\boldsymbol{\alpha}, M^*) = 0$. In that extreme case, $B(\boldsymbol{\alpha}, M) = 0$.

For any random vector $\boldsymbol{\alpha}$, define the semi-norm $|||\cdot|||_{\boldsymbol{\alpha}}$ induced by $\boldsymbol{\alpha}$ as:

$$|||A|||_{\boldsymbol{\alpha}} \triangleq \sum_{k=1}^{K} \mathbb{E}\Big[|\sum_{j=1}^{K} A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0)\Big].$$

Note that the subscript $\boldsymbol{\alpha}$ in $|||\cdot|||_{\boldsymbol{\alpha}}$ is used to indicate the dependence on the distribution of $\boldsymbol{\alpha}$. $|||\cdot|||_{\boldsymbol{\alpha}}$ is a semi-norm but not a norm because $|||A|||_{\boldsymbol{\alpha}} = 0$ does not imply $A = 0$. Actually, for any nonzero diagonal matrix $A \neq 0$, $|||A|||_{\boldsymbol{\alpha}} = 0$ because $\sum_{k=1}^{K} \mathbb{E}\Big[|A_{k,k}\boldsymbol{\alpha}_k|\mathbf{1}(\boldsymbol{\alpha}_k = 0)\Big] = 0$. Note that the reason why we define B and $|||\cdot|||_{\boldsymbol{\alpha}}$ this way is because these quantities appear naturally in the first order optimality condition of ℓ_1-minimization. Hopefully, the motivation of defining these definitions will become clear later.

Assumption I (Regular data-dependent norm) $|||\cdot|||_{\boldsymbol{\alpha}}$ is $c_{\boldsymbol{\alpha}}$-regular: There exists a number $c_{\boldsymbol{\alpha}} > 0$, dependent on the distribution of $\boldsymbol{\alpha}$, such that for any matrix $A \in H^K$, where $H^K = \{A \in \mathbb{R}^{K\times K} \mid A_{i,i} = 0 \text{ for all } 1 \leq i \leq K\}$, $|||A|||_{\boldsymbol{\alpha}}$ is bounded below by A's Frobenious norm: $|||A|||_{\boldsymbol{\alpha}} \geq c_{\boldsymbol{\alpha}}|||A|||_F$.

Note that similar to $|||\cdot|||_{\boldsymbol{\alpha}}$, we use the subscript $\boldsymbol{\alpha}$ in $c_{\boldsymbol{\alpha}}$ to indicate that the quantity $c_{\boldsymbol{\alpha}}$ depends on the distribution of $\boldsymbol{\alpha}$. Assumption I has several implications. First, it ensures that the coefficient vector $\boldsymbol{\alpha}$ does not lie in a linear subspace of $\mathbb{R}^K$. Otherwise, we can make rows of A orthogonal to $\boldsymbol{\alpha}$ and show that $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular. Second, it also guarantees that the coefficient vector $\boldsymbol{\alpha}$ must have some level of sparsity. To see why this is the case, suppose there exists some coordinate k' such that the coefficient $\boldsymbol{\alpha}_{k'} \neq 0$ almost surely. We can then construct A such that all of its elements are zero except the k'-th row. Thus, $|||A|||_{\boldsymbol{\alpha}} = \mathbb{E}\Big[|\sum_{j=1}^{K} A_{k',j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_{k'} = 0)\Big] = 0$, but $|||A|||_F > 0$. Third, if we define the dual (semi-)norm of $|||\cdot|||_{\boldsymbol{\alpha}}$ in the subspace H^K as

$$|||X|||_{\boldsymbol{\alpha}}^* = \sup_{A\neq 0, A\in H^K} \frac{\operatorname{tr}(X^T A)}{|||A|||_{\boldsymbol{\alpha}}}, \quad \text{for } X \in \mathbb{R}^{K\times K},$$

the regularity of $|||\cdot|||_{\boldsymbol{\alpha}}$ implies that the corresponding dual semi-norm is bounded above by the Frobenius norm. To see this, simply note that $|||X|||_{\boldsymbol{\alpha}}^* \leq \frac{1}{c_{\boldsymbol{\alpha}}}|||X|||_F$ with the above definition. Assumption I is crucial for the study of the local identifiability property. As can be seen later in Theorems 4.1.1 and 4.1.2, regularity of $|||\cdot|||_{\boldsymbol{\alpha}}$ is indispensable in determining the sharpness of the local minimum corresponding to the reference dictionary $\boldsymbol{D}^*$ as well as the bounding region.

Assumption II (Probabilistic linear independence) For any fixed constants $c_1, \ldots, c_K \in \mathbb{R}$, the following statement holds almost surely

$$\sum_{l=1}^{K} c_l\boldsymbol{\alpha}_l = 0 \Longrightarrow c_l\boldsymbol{\alpha}_l = 0 \ \forall\ l = 1, \ldots, K,$$

or equivalently, for any fixed $c_1, \ldots, c_K$,

$$P\left(\sum_{l=1}^{K} c_l\boldsymbol{\alpha}_l = 0, \sum_{l=1}^{K} c_l^2\boldsymbol{\alpha}_l^2 > 0\right) = 0.$$

Assumption II controls the sparsity of any coefficient vector $\boldsymbol{\beta}$ under a general dictionary $\boldsymbol{D}$. For the noiseless signal $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{\alpha}$, its j-th coefficient under a dictionary $\boldsymbol{D}$ can be written as a linear combination of reference coefficients $\boldsymbol{\alpha}_l$: $\boldsymbol{\beta}_j = \boldsymbol{D}^{-1}[j,]\boldsymbol{D}^*\boldsymbol{\alpha} = \sum_{l=1}^{K} c_l\boldsymbol{\alpha}_l$ where $c_l = \boldsymbol{D}^{-1}[j,]\boldsymbol{D}_l^*$ for $l = 1,\ldots,K$. Thus, Assumption II implies that under any general dictionary, the resulting coefficient $\boldsymbol{\beta}_j$ is zero if and only if for each l, either the reference coefficient is zero ($\boldsymbol{\alpha}_l = 0$) or the corresponding constant is zero ($c_l = 0$). In other words, elements in the reference coefficient vector cannot "cancel" with each other unless all the elements are zeros. This assumption seems very similar to the *linear independence* property of random variables [76]: Random variables $\psi_1,\ldots,\psi_K$ are linearly independent if $c_1\psi_1 + \ldots + c_K\psi_K = 0$ a.s. implies $c_1 = c_2 = \cdots = c_K = 0$. It is worth pointing out that Assumption II is a weaker assumption than linear independence. Many distributions of interest, such as Bernoulli Gaussian distributions, are not linearly independent but satisfy Assumption II (Proposition 3.2.2). This assumption is essential when we study the uniqueness of the sharp local minimum in Theorem 4.2.1.

In the following propositions, we show that both Bernoulli-type models and exact sparse models satisfy Assumption I and II.

Proposition 3.2.1. *The norm $|||\cdot|||_{\boldsymbol{\alpha}}$ induced by exact sparse models or Bernoulli-type models satisfy Assumption I. The regularity constant has explicit form when the coefficient is from $SG(s)$ or $BG(p)$:*

- *If $\boldsymbol{\alpha}$ is from $SG(s)$, the norm $|||\cdot|||_{\boldsymbol{\alpha}}$ is c_s-regular, where $c_s \geq \frac{s(K-s)}{K(K-1)}\sqrt{\frac{2}{\pi}}$.*
- *If $\boldsymbol{\alpha}$ is from $BG(p)$, the norm $|||\cdot|||_{\boldsymbol{\alpha}}$ is c_p-regular, where $c_p \geq p(1-p)\sqrt{\frac{2}{\pi}}$.*

Proposition 3.2.2. *If the coefficient vector is generated from a Bernoulli-type model or an exact sparse model, Assumption II holds.*

Remarks: Although the above assumptions are quite general, certain distributions considered in other papers do not satisfy our assumptions. A key requirement in Bernoulli-type or exact sparse models is that the probability density function of the base random variable z must exist. For instance, the Bernoulli Rademacher model [84] does not satisfy Assumption II. To see this, take the following Bernoulli Rademacher model for $K = 2$ as an example: suppose $\boldsymbol{\xi} \in \{0,1\}^2$ where $P(\xi_1 = 1) = p_1$, $P(\xi_2 = 1) = p_2$. The base random vector $z \in \{-1,1\}^2$ with $P(z_1 = 1) = P(z_2 = 1) = 1/2$. If we take $c_1 = 1$ and $c_2 = -1$, $P(c_1\boldsymbol{\alpha}_1 + c_2\boldsymbol{\alpha}_2 = 0, c_1\boldsymbol{\alpha}_1 \neq 0, c_2\boldsymbol{\alpha}_2 \neq 0) = P(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2 = 0, \xi_1 \neq 0, \xi_2 \neq 0) = P(\xi_1 = 1, \xi_2 = 1, z_1 = z_2) = p_1 \cdot p_2/2 > 0$. Therefore, Assumption II does not apply in this case.

Chapter 4

Local and global identifiability

Similar to [100], we first study the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{D}}{\text{minimize}} \quad & \mathbb{E}L(\boldsymbol{D}) = \mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \\ \text{subject to} \quad & \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \end{aligned} \tag{4.1}$$

Here, the notation $\mathbb{E}$ is the expectation with respect to $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{\alpha}$ under a probabilistic model for $\boldsymbol{\alpha}$. Therefore, this optimization problem is equivalent to the case when we have infinite number of samples. As we shall see, the analysis of this population level problem gives us significant insights into the identifiability properties of dictionary learning. We also consider the finite sample case (3.2) in Theorem 4.2.2.

4.1 A necessary and sufficient condition for local identifiability

In this subsection, we will establish a necessary and sufficient condition for the reference dictionary to be a sharp local minimum.

Theorem 4.1.1 (Local identifiability)**.** *Suppose $|||\cdot|||_{\boldsymbol{\alpha}}$ is $c_{\boldsymbol{\alpha}}$-regular (see Assumption I) and the ℓ_1 norm of the reference coefficient vector $\boldsymbol{\alpha}$ has bounded first order moment: $\mathbb{E}\|\boldsymbol{\alpha}\|_1 < \infty$. $\boldsymbol{D}^*$ is a sharp local minimum of Formulation (4.1) with sharpness at least $\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}^*|||_2^2}(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)$ if and only if*

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* < 1. \tag{4.2}$$

If $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* > 1$, $\boldsymbol{D}^*$ is not a local minimum.

Remarks: [100] studied the local identifiability problem when the coefficient vector $\boldsymbol{\alpha}$ is from Bernoulli Gaussian or sparse Gaussian distributions. They gave a sufficient and almost

necessary condition that ensures the reference dictionary to be a local minimum. Theorem 4.1.1 substantially extends their result in two aspects:

- The reference coefficient distribution can be exact sparse models and Bernoulli type models, which is more general than sparse/Bernoulli Gaussian models.
- In addition to showing that the reference dictionary $\boldsymbol{D}^*$ is a local minimum, we show that $\boldsymbol{D}^*$ is actually a sharp local minimum with an explicit bound on the sharpness.

To prove Theorem 4.1.1, we need to calculate how the objective function changes along any direction in the neighborhood of the reference dictionary. The major challenge of this calculation is that the objective function is neither convex nor smooth, which prevents us from using sub-gradient or gradient to characterize its local structure. Instead, we obtain a novel sandwich-type inequality of the ℓ_1 objective function (Lemma A.4.4). With the help of this inequality, we are able to carry out a more fine-grained analysis of the ℓ_1-minimization objective. The detailed proof of Theorem 4.1.1 can be found in the Appendix.

Theorem 4.1.1 gives the condition under which the reference dictionary is a sharp local minimum. The below Theorem 4.1.2 gives an explicit bound of the size of the region. To the best of authors' knowledge, this is the first result about the region where local identifiability holds for ℓ_1-minimization.

Theorem 4.1.2. *Under notations in Theorem 4.1.1, if* $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* < 1$, *for any* $\boldsymbol{D}$ *in the set*

$$S = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2|||\boldsymbol{D}^*|||_2, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*) \cdot c_{\boldsymbol{\alpha}}}{8\sqrt{2}|||\boldsymbol{D}^*|||_2^2 \max_j \mathbb{E}|\boldsymbol{\alpha}_j|} \right\},$$

we have $\mathbb{E}L(\boldsymbol{D}) \geq \mathbb{E}L(\boldsymbol{D}^*)$.

Remarks: First of all, note that the set S we study here is different from what [2] called the "basin of attraction". The basin of attraction of an iterative algorithm is the set of initialization dictionaries under which the algorithm converges to the reference dictionary $\boldsymbol{D}^*$. For an iterative algorithm that decreases its objective function at each step, its basin of attraction must be a subset of the region within which $\boldsymbol{D}^*$ has the minimal objective value. Second, Theorem 4.1.2 only tells us that $\boldsymbol{D}^*$ admits the smallest objective function value within the set S. It does not, however, indicate that $\boldsymbol{D}^*$ is the only local minimum within S.

For certain generative models, the conditions in Theorem 4.1.1 and 4.1.2 can be made more explicit to compare with other local identifiability results. In what follows, we will study two examples to gain a better understanding of those conditions. These examples demonstrate the trade-off between coefficient sparsity, collinearity of atoms in the reference dictionary and signal dimension K. For simplicity, we set the reference dictionary to be the

constant collinearity dictionary with coherence $\mu > 0$: $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu\mathbf{1}\mathbf{1}^T)^{1/2}$ where $\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{K\times K}$ is a square matrix whose elements are all ones. This simple dictionary class was used to illustrate the local identifiability conditions in [32] and [100]. The coherence parameter μ controls the collinearity between dictionary atoms. By studying this class of reference dictionaries, we can significantly simplify the conditions and demonstrate how the coherence μ affects dictionary identifiability.

Corollary 4.1.1. *Suppose the reference dictionary $\boldsymbol{D}^*$ is a constant collinearity dictionary with coherence $\mu > 0$: $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu\mathbf{1}\mathbf{1}^T)^{1/2}$, and the reference coefficient vector $\boldsymbol{\alpha}$ is from $SG(s)$. $\boldsymbol{D}^*$ is a sharp local minimum with sharpness at least $\frac{s}{\sqrt{\pi}(1+\mu(K-1))K}\left(\frac{K-s}{K-1} - \mu\sqrt{s}\right)$ if and only if*

$$\mu\sqrt{s} < \frac{K-s}{K-1}.$$

If the above inequality holds true, for any

$$\boldsymbol{D} \in S = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{\frac{K-s}{K-1} - \mu\sqrt{s}}{8\sqrt{2}(1+\mu(K-1))} \right\},$$

we have $\mathbb{E}L(\boldsymbol{D}) \geq \mathbb{E}L(\boldsymbol{D}^*)$.

Three parameters play important roles for the reference dictionary to be a sharp local minimum: dictionary coherence μ, sparsity s and dimension K. Since $\mu\sqrt{s} - \frac{K-s}{K-1}$ is a monotonically increasing function with respect to μ and s, local identifiability holds when the dictionary is close to an orthogonal matrix and the coefficient vector is sufficiently sparse. Another important observation is that $\mu\sqrt{s} - \frac{K-s}{K-1}$ is monotonically decreasing as K increases. Thus, given that the number of nonzero elements per signal s is fixed, it is easier for the local identifiability condition to hold for larger K. If K tends to infinity, the condition becomes $s < \frac{1}{\sqrt{\mu}}$. Also, the set S shrinks as s or μ increases, implying that the region is smaller when the coefficients are less sparse or the dictionary has higher coherence. When $\mu = 0$, the set S becomes $\left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{1}{8\sqrt{2}}\frac{K-s}{K-1} \right\}$.

Next, we consider non-negative sparse Gaussian distribution in the following example. Since we do not have the explicit form of the regularity constant $c_{\boldsymbol{\alpha}}$ for non-negative sparse Gaussian distribution, we omit the corresponding results for the sharpness and the region bound.

Corollary 4.1.2. *Suppose the reference dictionary is a constant collinearity dictionary with coherence $\mu > 0$: $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu\mathbf{1}\mathbf{1}^T)^{1/2}$, and the reference coefficient vector $\boldsymbol{\alpha}$ is from non-negative sparse Gaussian distribution $|SG(s)|$. If*

$$\left|\mu - \frac{s-1}{K-1}\right| < \frac{K-s}{K-1},$$

then $\boldsymbol{D}^*$ is a sharp local minimum.

Note that the condition $\frac{K-1}{K-s} \cdot \left|\mu - \frac{s-1}{K-1}\right| < 1$ is equivalent to $\frac{2s-K-1}{K-1} < \mu < 1$. When K tends to infinity, the reference dictionary is a local minimum for $\mu < 1$. Compared to the same bound from Corollary 4.1.1, $\mu < \frac{1}{\sqrt{s}}$ for large K, the bound for non-negative coefficients is less restrictive. Therefore, the non-negativity of the coefficient distribution relaxes the requirement for local identifiability.

Corollary 4.1.3. *Let the reference dictionary be a constant collinearity dictionary with coherence μ. Assume that the reference coefficients are generated from the Bernoulli Gaussian model $BG(p)$. If $\frac{\mu\sqrt{p(K-1)}}{1-p} < 1$, the reference dictionary is a sharp local minimum of $\mathbb{E}L(D)$ with sharpness at least $\frac{p}{\sqrt{\pi}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right)$. In addition, for any*

$$\boldsymbol{D} \in \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, \right.$$
$$\left. |||\boldsymbol{D}-\boldsymbol{D}^*|||_F^2 \leq \frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right) \right\},$$

we have $\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \mathbb{E}\|\boldsymbol{\alpha}\|_1$.

Corollary 4.1.4. *Let the reference dictionary be a constant collinearity dictionary with coherence μ. Assume that the coefficients are generated from the sparse Laplacian model $SL(s)$. If*

$$\frac{\mu s(K-1)}{(K-s)\int\int_0^\infty |y-x|(xy)^{s-1}\exp(-(x+y))\Gamma(s)^{-2}dxdy} < 1,$$

then the reference dictionary is a sharp local minimum of $\boldsymbol{D} \mapsto \mathbb{E}L(\boldsymbol{D})$.

Although the condition in Corollary 4.1.4 is quite convoluted, we can compare it with the sparse Gaussian case empirically. For sparse Gaussian distributions, there are two parameters: sparsity s and dimension K. Define the phase transition curve to be the asymptotic boundary that separates the region where local identifiability holds (the area below the curve) and the region where local identifiability fails (the area above the curve). When $K = 10$ and $K = 20$, the phase transition curve ($|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = 1$) for sparse Laplace distribution and sparse Gaussian distribution can be found in Fig. 4.1. As can be seen in the figure, the phase transition curve for sparse Laplace distribution is slightly higher than that for sparse Gaussian distribution, suggesting that the Laplace distribution has less stringent local identifiability condition. That is consistent with our intuition: while the density function of a standard Gaussian distribution is rotation symmetric, which implies that it does not prefer any direction, the density function of the Laplace distribution is not. For example, consider a simple two-dimensional case: let $\boldsymbol{D}^*$ be the identity matrix in $\mathbb{R}^{2\times 2}$. If the reference coefficient is from the standard Gaussian distribution with no sparsity, i.e. $s = K$, all the orthogonal dictionaries will have the same objective value $\sqrt{\frac{2}{\pi}}$. So local identifiability does not hold for Gaussian distribution under the setting $s = K$. However, for the Laplace distribution, even

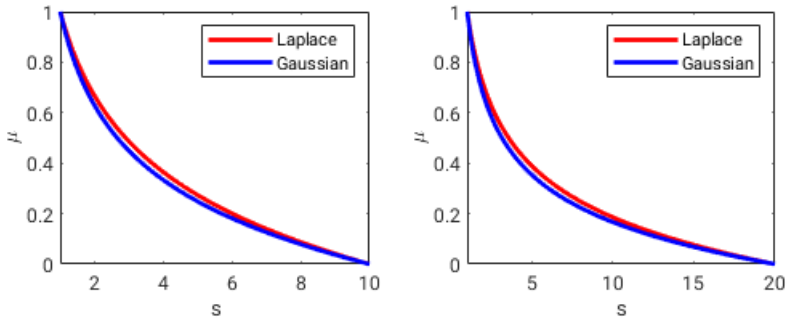


Figure 4.1: Theoretical phase transition curve for constant collinearity dictionary with coherence μ and sparsity s for $K = 10$ (left) and $K = 20$ (right).

if $s = K$, for an orthogonal dictionary $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ with $\theta \in [0, \pi/2]$, its ℓ_1 objective function value is $2(\sin\theta + \cos\theta + \frac{1}{\sin\theta+\cos\theta})$, which attains its minimum when $\theta = 0$ or $\frac{\pi}{2}$. This means the local identifiability still holds.

4.2 A counterexample of global identifiability and a remedy

For ℓ_1-minimization, multiple local minima exist: as a result of sign-permutation ambiguity, if $\boldsymbol{D}$ is a local minimum, for any permutation matrix P and any diagonal matrix Λ with diagonal elements ± 1, $\boldsymbol{D}P\Lambda$ is also a local minimum. These local minima are benign in nature since they essentially refer to the same dictionary. Can there be local minima other than the benign ones? If so, how can we distinguish benign local minima from them? In this subsection, we consider the problem of global identifiability. First, we give a counterexample to show that the reference dictionary is not necessary a global minimum of the ℓ_1-minimization problem even for orthogonal dictionary and sparse coefficients.

Counterexample on global identifiability. Suppose the reference dictionary is the identity matrix $\mathbb{I} \in \mathbb{R}^{2\times 2}$. The coefficients are generated from a Bernoulli-type model $\boldsymbol{\alpha} \in \mathbb{R}^2$ such that $\boldsymbol{\alpha}_i = z_i\xi_i$ for $i = 1, 2$, where ξ_1 and ξ_2 are Bernoulli variables with success probability 0.67, and (z_1, z_2) is drawn from the below Gaussian mixture model:

$$\frac{1}{2}\mathcal{N}\left(0, \begin{pmatrix} 101 & -99 \\ -99 & 101 \end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(0, \begin{pmatrix} 101 & 99 \\ 99 & 101 \end{pmatrix}\right).$$

We generate 2000 samples from the model and compute the dictionary learning objective $L(\cdot)$ defined in (3.1) for each candidate dictionary (Fig. 4.2). As can be seen from the

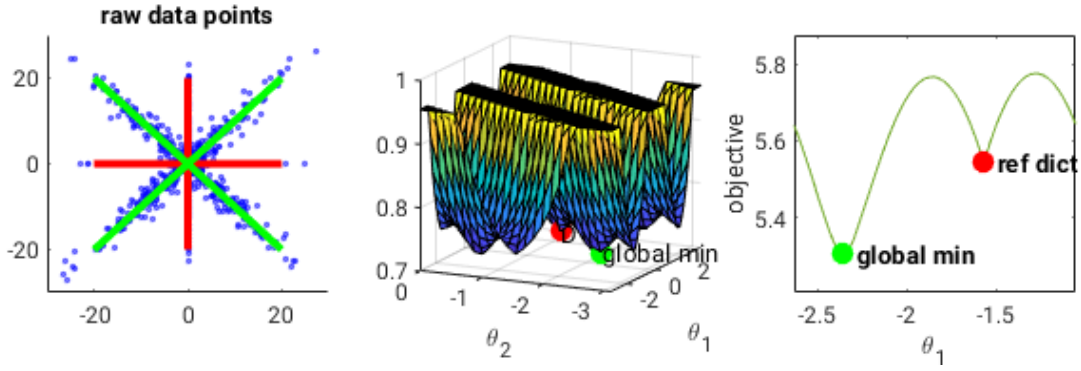


Figure 4.2: The empirical data (left) and the objective surface plot (middle). We parameterize a candidate dictionary as $\boldsymbol{D} = (a_1, a_2)$, where $a_1 = (\cos(\theta_1), \sin(\theta_1))$, $a_2 = (\cos(\theta_2), \sin(\theta_2))$. The objective of $\boldsymbol{D}$ is defined as in (3.1). Green dots/lines indicate global minima, whereas red dots/lines are the reference dictionary or its sign-permutation equivalents. The right figure shows the objective curve for all orthogonal dictionaries ($\theta_1 - \theta_2 = \pi/2$). While the reference dictionary is a sharp local minimum, it is not a global minimum.

objective function surface plot, the global minimum for this data set is not the reference dictionary. Furthermore, one can show that the global minimum is not sharp, i.e., the directional derivative along certain directions at the global minimum is close to zero for finite samples and exactly zero for the population case.

The above example shows a potential drawback of directly minimizing the ℓ_1 objective compared to other objectives such as ℓ_0. For the ℓ_0 objective, under certain Bernoulli Gaussian models, the reference dictionary is a global minimum [84] and even in our counterexample, which is not Bernoulli Gaussian, the reference dictionary can still be shown to be a global minimum. Still, the computation complexity of the ℓ_0 objective remains too high to switch from ℓ_1. To remedy this drawback of ℓ_1, we observe that in the above example, although the reference dictionary is not a global minimum, it is still a sharp local minimum and there are no other sharp local minima. Therefore there is hope that we can combine the ℓ_1 objective and a "sharpness" test to recover the reference dictionary. Is this observation true for general cases? The answer is yes. The following theorem shows that the reference dictionary is the unique sharp local minimum of ℓ_1-minimization up to sign-permutation.

Theorem 4.2.1 (Unique sharp local minimum)**.** *Suppose the reference coefficient vector $\boldsymbol{\alpha}$ satisfies probabilistic linear independence (see Assumption II). If $\boldsymbol{D}^*$ is a sharp local minimum of Formulation (4.1), it is the only sharp local minimum in $\mathbb{B}(\mathbb{R}^K)$. If it is not a sharp local minimum, there are no sharp local minima in $\mathbb{B}(\mathbb{R}^K)$.*

Note that Theorem 4.2.1 works for the population case where the sample size is infinite. For the finite sample case, we can show that the sharpness of spurious local minima is close

to zero. Define $\mathcal{D}_\epsilon$ to be

$$\mathcal{D}_\epsilon = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| \boldsymbol{D} \text{ is a sharp local minimum of (3.2) with sharpness at least } \epsilon. \right\}.$$

Define $\text{eig}(\boldsymbol{D})$ to be the set of eigenvalues of the matrix $\boldsymbol{D}$. For any fixed $\epsilon > 0$ and $\rho_2 > \rho_1 > 0$, define the event $A(\rho_1, \rho_2, \epsilon)$ to be

$$\begin{aligned} A(\rho_1, \rho_2, \epsilon) = \{&\text{There exists } \boldsymbol{D} \in \mathcal{D}_\epsilon \text{ s.t. } \text{eig}(\boldsymbol{D}) \subset (\rho_1, \rho_2) \text{ and} \\ &\boldsymbol{D} \neq \boldsymbol{D}^* \text{ up to sign-permutation.}\}. \end{aligned} \tag{4.3}$$

In other words, the event A represents the "bad" event that at least one of the sharp local minima in $\mathcal{D}_\epsilon$ with bounded eigenvalues is not the reference dictionary. With this notation, Theorem 4.2.1 basically shows that for the population case, the event $A(\rho_1 = 0, \rho_2 = \infty, \epsilon = 0)$ will never happen. The next theorem shows that for the finite sample case, $P(A(\rho_1, \rho_2, \epsilon))$ is upper bounded.

Theorem 4.2.2 (Finite-sample case)**.** *Suppose $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$ are drawn i.i.d. from a model satisfying probabilistic linear independence (see Assumption II) and for any $i = 1 \ldots n$, $\|\boldsymbol{x}^{(i)}\|_2 \leq L < \infty$. Then for any fixed $\rho_2 > \rho_1 > 0$ and $\epsilon > 0$,*

$$P\left(A(\rho_1, \rho_2, \epsilon)\right) \leq 4 \exp\left(2K \left(\ln \frac{n}{2K} + 1 \right) - n \left(\frac{\rho_1^3 \epsilon}{2L\rho_2} - \frac{1}{n} \right)^2 \right).$$

In particular, $P(A(\rho_1, \rho_2, \epsilon)) \to 0$ as $\frac{K}{n} \to 0$.

Remarks: Theorem 4.2.2 ensures that as $K/n \to 0$, with high probability *no* dictionaries other than $\boldsymbol{D}^*$ are sharp local minima within a region $\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \big| \text{eig}(\boldsymbol{D}) \in (\rho_1, \rho_2)\}$. However, it does not tell whether or not $\boldsymbol{D}^*$ is a sharp local minimum. For the population case, this issue is resolved in Theorem 4.1.1, which gives a necessary and sufficient condition for the reference dictionary to be a sharp local minimum.

Chapter 5

Algorithms for checking sharpness and solving ℓ_1-minimization

As shown in the previous section, the reference dictionary is the unique sharp local minimum under certain conditions. Here, we will design an algorithm that uses this property as a stopping criterion for ℓ_1-minimization. If the algorithm finds a sharp local minimum, we know that it is the reference dictionary. To do so we need to answer the following practical questions:

- How to determine numerically if a given dictionary is a sharp local minimum?
- How to find a sharp local minimum and recover the reference dictionary?

In this section, we first introduce an algorithm to check if a given dictionary is a sharp local minimum. We then develop an algorithm to recover the reference dictionary. The latter algorithm is guaranteed to decrease the (truncated) ℓ_1 objective function at each iteration (Proposition 5.2.2).

5.1 Determining sharp local minima

Despite the intuitive concept, checking whether a given dictionary is a sharp local minimum can be challenging. First of all, the dimension of the problem is very high (p^2). Second, if a dictionary is a sharp local minimum, the objective function is not differentiable at that point, precluding us from using gradients or the Hessian to solve the problem. One might also consider using sub-gradients to minimize the objective [8]. However, because the problem is actually non-convex, sub-gradients might not be well-defined.

We propose a novel algorithm to address these challenges. We decompose the problem into a series of sub-problems each of which is low-dimensional. In Proposition 5.1.1, we show that a given dictionary is a sharp local minimum in dimension p^2 if and only if certain vectors are sharp local minima for the corresponding sub-problems of dimension p. The objective

function of each subproblem is strongly convex. To deal with non-existence of gradient or Hessian, we design a perturbation test based on the observation that a sharp local minimum ought to be stable with respect to small perturbations. For instance, $x = 0$ is the sharp local minimum of $|x|$ but is non-sharp local minimum of x^2. If we add a linear function as a perturbation, $x = 0$ is still a local minimum of $|x| + \epsilon \cdot x$ for any ϵ such that $|\epsilon| < 1$ but not so for $x^2 + \epsilon \cdot x$. The choice of the perturbation is crucial. In Proposition 5.1.1, we show that adding a perturbation to the dictionary collinearity matrix M is sufficient. Note that perturbations to other quantities might work as well. Intuitively, a "good" perturbation should provide enough variability along any direction. Otherwise, a local minimum that is not sharp along certain directions might be mistakenly deemed as sharp.

Proposition 5.1.1. *The following three statements are equivalent:*

1) $\boldsymbol{D}$ *is a sharp local minimum of* (3.2).

2) For any $k = 1, \ldots, p$, $\mathbb{I}_k$ *is the sharp local minimum of the strongly convex optimization:*

$$\mathbb{I}_k \in \operatorname{argmin}_{\boldsymbol{w}} \mathbb{E}|\langle \boldsymbol{\beta}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{p} \sqrt{(w_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \mathbb{E}|\boldsymbol{\beta}_h|. \tag{5.1}$$
$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_p] \in \mathbb{R}^p, \ w_k = 1.$$

3) For a sufficiently small $\rho > 0$ *and any* $\tilde{M}$ *s.t.* $|\tilde{M}_{k,h} - M_{k,h}| \leq \rho$ *for any* $k, h = 1, \ldots, p$, $\mathbb{I}_k$ *is the local minimum of the convex optimization:*

$$\mathbb{I}_k \in \operatorname{argmin}_{\boldsymbol{w}} \mathbb{E}|\langle \boldsymbol{\beta}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{p} \sqrt{(w_h - \tilde{M}_{k,h})^2 + 1 - \tilde{M}_{k,h}^2} \cdot \mathbb{E}|\boldsymbol{\beta}_h|. \tag{5.2}$$
$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_p] \in \mathbb{R}^p, \ w_k = 1.$$

for $k = 1, \ldots, p$.

Proposition 5.1.1 tells us that, in order to check whether a dictionary is a sharp local minimum, it is sufficient to add a perturbation to the matrix $M = \boldsymbol{D}^T\boldsymbol{D}$ and check whether the resulting dictionary is the local minimum of the perturbed objective function. Empirically, we can add a Gaussian noise with a small enough variance ρ and minimize the objective (5.2). If $\mathbb{I}_k$, the k-th column vector of the identity matrix, is the local minimum for the perturbed objective, by Proposition 5.1.1 the given dictionary is guaranteed to be a sharp local minimum. We formalize this idea into Algorithm 1. We acknowledge that this algorithm might be conservative and misclassify a sharp local minimum as a non-sharp local minimum if ρ is not small enough as required in Proposition 5.1.1. There is no good rule-of-thumb in choosing ρ as it can be dependent on the data. We will explore the sensitivity of this algorithm with respect to choice of ρ in the simulation section.

Algorithm 1 Sharp local minimum test for ℓ_1-minimization dictionary learning

Require: Dictionary to be tested $\boldsymbol{D}$, samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$, perturbation level $\rho \in \mathbb{R}^+$, threshold $T \in \mathbb{R}^+$.

for $i = 1, \dots, n$ **do**

$\boldsymbol{\beta}^{(i)} \leftarrow \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$.

end for

for $j = 1, \dots, p$ **do**

Generate $\epsilon_j \sim \mathcal{N}(0, \rho \cdot \mathbb{I}_{p\times p})$.

$\tilde{\boldsymbol{D}}_j = \boldsymbol{D}_j + \epsilon_j$.

end for

for $k, h = 1, \dots, p$ **do**

$\tilde{M}_{k,h} \leftarrow \left\langle \tilde{\boldsymbol{D}}_h, \tilde{\boldsymbol{D}}_k \right\rangle$ if $k \neq h$ or 0.

end for

$r \leftarrow 0$

for $k = 1, \dots, p$ **do**

Solve the strongly convex optimization via BFGS:

$$\boldsymbol{w}^{(k)} \leftarrow \text{minimize}_{\boldsymbol{w}} \sum_{i=1}^{n} |\langle \boldsymbol{\beta}^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{p} \sqrt{(w_h - \tilde{M}_{k,h})^2 + 1 - \tilde{M}_{k,h}^2} \cdot \sum_{i=1}^{n} |\boldsymbol{\beta}_h^{(i)}|. \tag{5.3}$$

$$\text{subject to } \boldsymbol{w} = [w_1, \dots, w_p] \in \mathbb{R}^p, \quad w_k = 1. \tag{5.4}$$

$\mathbb{I}_k \leftarrow (0, \dots, 0, 1, 0, \dots, 0)$ where only the k-th element is 1.

$r \leftarrow \max(r, \|\boldsymbol{w}^{(k)} - \mathbb{I}_k\|_2^2)$.

end for

if $r < T$ **then**

Output $\boldsymbol{D}$ is a sharp local minimum.

else

Output $\boldsymbol{D}$ is not a sharp local minimum.

end if

The main component of Algorithm 1 is solving the strongly convex optimization (5.3). To do so we use Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [97], which is a second order method that estimates Hessian matrices using past gradient information. Each step of BFGS is of complexity $O(np + p^2)$. If we assume the maximum iteration to be a constant, the overall complexity of Algorithm 1 is $O(np^2 + p^3)$. Because sample size n is usually larger than the dimension p, the dominant term in the complexity is $O(np^2)$. In the simulation section, we show that the empirical computation time is in line with the theoretical bound.

5.2 Recovering the reference dictionary

We now try to solve formulation (3.2). One of the most commonly used technique in solving dictionary learning is alternating minimization [66, 59], which is to update the coefficients and the dictionary in an alternating fashion until convergence. This method fails for noiseless ℓ_1-minimization: when the coefficients are fixed, the dictionary must also be fixed to satisfy all constraints. To allow dictionaries to be updated iteratively, researchers have proposed different ways to relax the constraints [2, 60]. However, those workarounds tend to have numerical stability issues if a high precision result is desired [60].

This motivates us to propose Algorithm 2. The algorithm uses the idea from Block Coordinate Descent (BCD). It updates each row of $\boldsymbol{D}^{-1}$ and the corresponding row in the coefficient matrix simultaneously. As we update one row of $\boldsymbol{D}^{-1}$, we also scale all the other rows of $\boldsymbol{D}^{-1}$ by appropriate constants. This is because if we only update one row of $\boldsymbol{D}^{-1}$ while keeping the others fixed, columns of the resulting dictionary will not have unit norm. The following lemma gives an admissible parameterization for updating one row of $\boldsymbol{D}^{-1}$.

Proposition 5.2.1. *For any dictionary $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^p)$ and any coordinate $k \in 1, \ldots, p$, given a vector $w = [w_1, \ldots, w_p] \in \mathbb{R}^p$ such that $w_k = 1$, we can define a matrix $Q \in \mathbb{R}^{p\times p}$:*

$$Q[k,] = \begin{cases} w^T \boldsymbol{D}^{-1} & h = k \\ \sqrt{(w_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \boldsymbol{D}^{-1}[h,] & h \neq k \end{cases}.$$

Then $Q^{-1} \in \mathbb{B}(\mathbb{R}^p)$, which means each column of Q^{-1} is of norm 1.

With the parameterization in Proposition 5.2.1, we derive the following subproblems from ℓ_1-minimization dictionary learning: for $k = 1, ..., K$,

$$\begin{aligned} &\text{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} |\langle \beta^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{p} \sqrt{(\boldsymbol{w}_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \sum |\beta_h^{(i)}|. \\ &\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_p] \in \mathbb{R}^p, w_k = 1. \end{aligned}$$

where $\boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$ for a dictionary $\boldsymbol{D}$. This new sub-problem is strongly convex, making it relatively easy to solve. Note that this problem is exactly the same as (5.1) in Proposition 5.1.1. Thus the optimization problem (5.1) is closely related to ℓ_1-minimization dictionary learning from two different perspectives: First, the sharpness of any solution of ℓ_1-minimization is equivalent to the sharpness of $\mathbb{I}_k$ for the optimization (5.1). Second, the optimization problem (5.1) can be viewed as a subproblem of ℓ_1-minimization under an appropriate parameterization.

A natural way to solve ℓ_1-minimization dictionary learning is to solve the above subproblems iteratively for each coordinate k. Similar ideas of learning a dictionary from a series of convex programs have been explored in other papers. For example, [84] reformulated the dictionary learning problem as a series of linear programs (LP) and construct a dictionary

from the LP solutions. Nonetheless, their algorithm is not guaranteed to minimize the ℓ_1 objective at each iteration.

We propose a coordinate-descent-based dictionary learning Algorithm 2. It has a tuning parameter τ, which aims at improving the performance of ℓ_1-minimization under the high signal-to-noise ratio settings. When τ is set to be infinity, Algorithm 2 minimizes the ℓ_1 objective at each update. However, when the signal-to-noise ratio is high, ℓ_1-minimization sometimes ends up with a low quality result. This is commonly due to the fact that the ℓ_1-norm over-penalizes large coefficients, which breaks the local identifiability, i.e., the reference dictionary is no longer a local minimum. Similar ideas are used in the re-weighted ℓ_1 algorithms in the field of compressed sensing [15]. The motivation of re-weighted algorithms is to reduce the bias of ℓ_1-minimization by imposing smaller penalty to large coefficients. In our algorithm, we simply truncate coefficient entries beyond the given threshold τ. The obtained problem is still strongly convex but this trick improves the numerical performance significantly.

The following theorem guarantees that the proposed algorithm always decreases the objective function value.

Proposition 5.2.2 (Monotonicity)**.** *Define*

$$f(\boldsymbol{D}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \min\left(\left|\boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)}\right|, \tau\right),$$

where τ is the threshold used in Algorithm 2. Denote by $\boldsymbol{D}^{(t,p)}$ the dictionary at the t-th iteration from Algorithm 2. $f(\boldsymbol{D}^{(t,p)})$ decreases monotonically for $t \in \mathbb{N}$: $f(\boldsymbol{D}^{(0,p)}) \geq f(\boldsymbol{D}^{(1,p)}) \geq f(\boldsymbol{D}^{(2,p)}) \ldots$

5.3 Numerical experiments

In this section, we evaluate the proposed algorithms with numerical simulations. We will study the empirical running time of Algorithm 1 in the first experiment and examine how the perturbation parameter ρ affects its performance in the second. In the third experiment, we study the sample size requirement for successful recovery of the reference dictionary. Finally, we will compare Algorithm 2 against other state-of-the-art dictionary learning algorithms [68, 69, 67]. The first two less computationally intensive simulations are run on an OpenSuSE OS with Intel(R) Core(TM) i5-5200U CPU 2.20GHz with 12GB memory, while the last two simulations are conducted in a cluster with 20 cores. The source code of the DL-BCD algorithm can be found in the github repository[1].

[1] https://github.com/shifwang/dl-bcd

Algorithm 2 Dictionary Learning Block Coordinate Descent (DL-BCD)

Require: Data $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$, threshold τ.
Initialize $\boldsymbol{D}^{(0,1)}$, $t \leftarrow 0$. $Q \leftarrow (\boldsymbol{D}^{(0,1)})^{-1}$.
while Stopping criterion not satisfied **do**
for $j = 1, \dots, p$ **do**
for $i = 1, \dots, n$ **do**
$\boldsymbol{\beta}^{(i)} \leftarrow Q\boldsymbol{x}^{(i)}$.
end for
for $h = 1, \dots, p$ **do**
$m_h \leftarrow \left\langle \boldsymbol{D}_h^{(t,j)}, \boldsymbol{D}_j^{(t,j)} \right\rangle$.
end for
Solve the convex optimization via BFGS:

$$\text{minimize}_{\boldsymbol{w}} \sum_{\substack{i=1..n,\\ |\beta_j^{(i)}|<\tau}} |\langle \boldsymbol{\beta}^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq j}^{p} \sqrt{(w_h - m_h)^2 + 1 - m_h^2} \cdot \sum_{\substack{i=1..n,\\ |\beta_h^{(i)}|<\tau}} |\beta_h^{(i)}|.$$

$$\text{subject to } \boldsymbol{w} = [w_1, \dots, w_p] \in \mathbb{R}^p, \ w_j = 1.$$

Update j-th row of Q: $Q[j,] \leftarrow \boldsymbol{w}^T Q$.
for $h = 1, \dots, p$, $h \neq j$ **do**
$Q[h,] \leftarrow Q[h,] \cdot \sqrt{(w_h - m_h)^2 + 1 - m_h^2}$.
end for
if $j = p$ **then**
$\boldsymbol{D}^{(t+1,1)} \leftarrow Q^{-1}$.
else
$\boldsymbol{D}^{(t,j+1)} \leftarrow Q^{-1}$.
end if
end for
$t \leftarrow t + 1$.
end while

Empirical running time of Algorithm 1

We evaluate the empirical computation complexity of Algorithm 1. Let the reference dictionary be a constant collinearity dictionary with coherence $\mu = 0.5$, i.e.,

$$\boldsymbol{D}^* = (0.5\mathbb{I} + 0.5\mathbf{1}\mathbf{1}^T)^{1/2},$$

The sparse linear coefficients are generated from the Bernoulli Gaussian distribution $BG(p)$ with $p = 0.7$. This specific parameter setting ensures that the reference dictionary is not a local minimum, thus making Algorithm 1 converge slower. For a fixed dimension, the

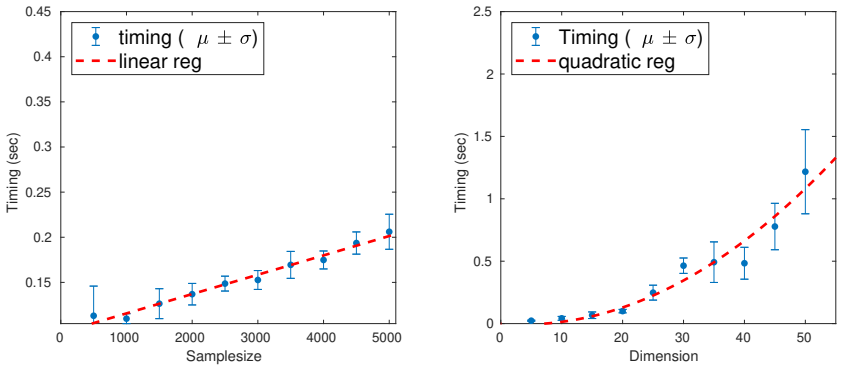


Figure 5.1: Computation time of Algorithm 1. Left: For $p = 20$ and $n = 500, \ldots, 5000$. Right: For $p = 5, \ldots, 50$ and $n = 400$.

computation time scales roughly linearly with the sample size, while for fixed sample size, the computation time scales quadratically with dimension p (Fig. 5.1). This shows that the empirical computation complexity of Algorithm 1 is of order $O(np^2)$, which is consistent with the theoretical complexity. Simulation results remain stable for different parameter settings.

Sensitivity analysis of the perturbation parameter ρ

In this experiment, we test the sensitivity of Algorithm 1 by varying the perturbation parameter ρ. We set dictionary dimension $p = 20$, sparsity parameter $s = 10$ and sample size $n = 1600$. Also, we consider constant collinearity dictionaries with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{p-s}{p-1}+0.1)$ (Fig. 5.2 Left) and $\mu = \frac{1}{\sqrt{s}}(\frac{p-s}{p-1}-0.2)$ (Fig. 5.2 Right). For the first experiment, the reference dictionary is not a sharp local minimum of the objective function given sufficiently large sample size. Hence a small perturbation to the dictionary results in a large distance between the global minimum of the perturbed optimization and $\mathbb{I}_k$, i.e., the quantity r defined in Algorithm 1. In the second experiment, the reference dictionary is sharp, indicating the distance r in Algorithm 1 should be small after adding a perturbation. For each value of ρ between 0.05 and 0.5, we repeat the algorithm 20 times to compute the resulting distances. When ρ is small, the distance r for the non-sharp case is very big (around 1.0) whereas for the sharp case it remains small (around 10^{-12}). For the sharp case, once ρ increases beyond 0.35, r increases drastically to 10^{-3}. This experiment shows for a wide range of parameter ρ values (0.05 to 0.3), Algorithm 1 succeeds in distinguishing between the sharp and non-sharp local minima. Nonetheless, there are two caveats when using this algorithm. First, the parameter ρ depends on the data generation process, which is usually not known in practice. Thus, it is still an open question about how to select ρ. Second, this algorithm is only useful for the noiseless case or when the noise is negligible. When the noise is significant, the reference dictionary is no longer a sharp local minimum. In that case, instead of checking

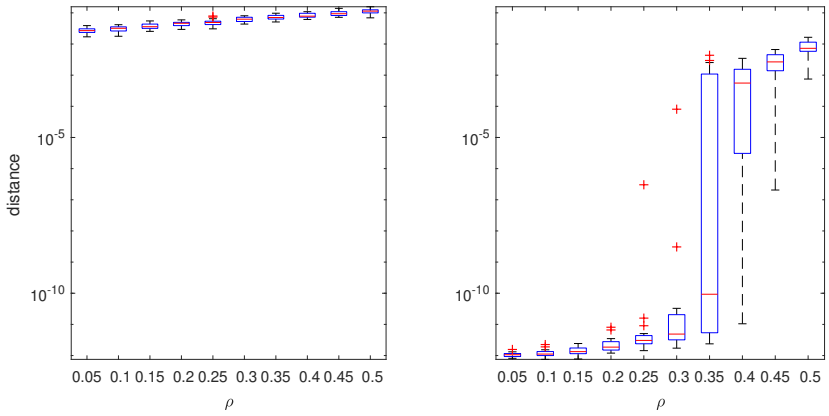


Figure 5.2: Sensitivity analysis of perturbation parameter ρ in Algorithm 1. Left: constant collinearity dictionary with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{p-s}{p-1} + 0.1)$; Right: constant collinearity dictionary with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{p-s}{p-1} - 0.2)$.

the sharpness, an alternative is to check the smallest eigenvalue of the Hessian. This idea has not been fully explored and will be studied in future work.

Empirical sample size requirement for local identifiability

When the reference dictionary is the constant collinearity matrix and the coefficients are sparse Gaussian, [100] shows that if the sample size n is of order $O(p \ln p)$, local identifiability holds with high probability. However, the corresponding constants that ensure local identifiability are unknown. In this subsection, we study the empirical sample size required for local identifiability with the help of Algorithm 1.

Suppose the reference dictionary has constant coherence $\mu = 0.5$ for various sizes $K = 12, 16, 20$ and the coefficients are drawn from the Sparse Gaussian distribution with sparsity $s = 5$. This specific parameter setting ensures the reference dictionary is a sharp local minimum given sufficient samples. Perturbation level is set at $\rho = 0.01$ and the threshold $T = 10^{-6}$. The experiment is repeated 20 times. Fig. 5.3 shows the percentage of experiments in which Algorithm 1 identifies $\boldsymbol{D}^*$ as a sharp local minimum for a variety of sample sizes n. Under this specific setting, to ensure local identifiability with 50 percent probability, the sample size n is roughly $20p$.

To further explore how dimension p affects the sample size for local identifiability, we run simulations for $p = 25, ..., 70$ and estimate the sample sizes that ensure the local identifiability with at least 50% chance. As shown in Fig. 5.4, the required sample size and dimension closely follow a linear relation $16.5p + 63$. It is linear, i.e., $O(p)$, instead of $O(p \ln p)$ because the sample size only ensures local identifiability with 50% chance.

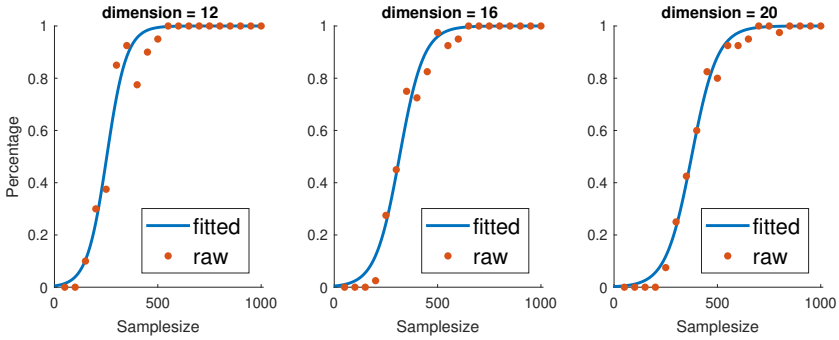


Figure 5.3: The percentage of experiments in which the reference dictionary is a local minimum, for different dimensions $p = 12, 16, 20$ and different sample sizes. The fitted line is obtained using a logistic regression. The sample size ensuring 50% chance is 253, 316, 375 respectively for $p = 12, 16, 20$.

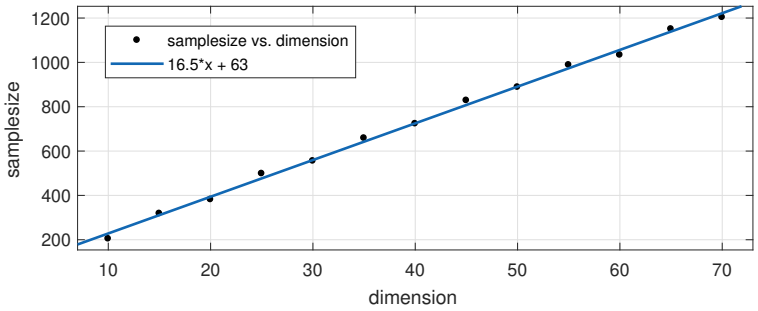


Figure 5.4: The estimated sample size that achieves 50 percent chance to ensure local identifiability for different p when the reference coefficient is generated from sparse Gaussian distribution and the reference dictionary has constant collinearity.

Global recovery performance of DL-BCD

The empirical performance of DL-BCD depends on the hyper-parameter τ and the initialization. In practice, we can set $\tau = \infty$ and initialize from a random orthogonal dictionary to obtain a rough approximation. Next, using the resulting dictionary as the initialization, we run the algorithm again with a reduced value of τ, say, 0.5. Compared with starting from a random initialization and using $\tau = 0.5$ directly, this two-step procedure performs better in our simulations.

In the below experiment, we set the reference dictionary to be a constant collinearity dictionary with $p = 10$ and $\mu = 0.7$, the coefficient is generated from sparse Gaussian model with sparsity $s = 4$ and the sample size is 400. Then we record the number of iterations for

DL-BCD to find the reference dictionary. As shown in Fig. 5.3, in 70 out of 100 experiments, DL-BCD recovers the reference dictionary with a single run. Note that if random initialization is used instead of ℓ_1-minimization with $\tau = \infty$, the one iteration recovery rate will drop from 70 percent to 5 percent. That means the initialization really plays an important role here: using ℓ_1-minimization as the initialization for truncated ℓ_1-minimization yields a good performance.

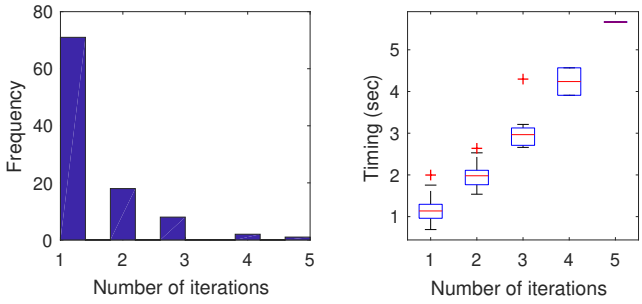


Figure 5.5: Histogram of number of iterations in DL-BCD and boxplot of the time complexity when initializing with the solution of ℓ_1-minimization with $\tau = \infty$.

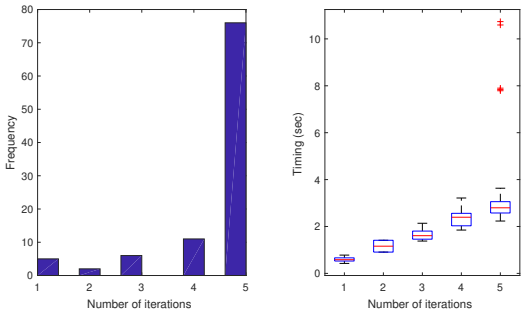


Figure 5.6: Histogram of number of iterations in DL-BCD and boxplot of the time complexity when initializing with random initialization. This shows random initialization is much worse than initializing with the solution of ℓ_1-minimization with $\tau = \infty$.

Comparison with other algorithms

We compare the performance of DL-BCD with other state-of-the-art algorithms, including the greedy K-SVD algorithm [3], SPAMS for online dictionary learning [61, 59], ER-SpUD(proj) for square dictionaries [84], and EM-BiG-AMP algorithm [68, 69]. The implementation of these algorithms is available in the package BiG-AMP [68, 69].

We use the following hyperparameter settings for each algorithm.

- EM-BiG-AMP: The outer loop that performs EM iterations is allowed up to 20 iterations. The inner loop is allowed a minimum of 30 and a maximum of 1500 iterations.
- K-SVD: K-SVD has two parameters: number of iterations and the enforced sparsity. The number of iterations is set to be 1000. The enforced sparsity is set to be the same as the true sparsity of the underlying model s.
- SPAMS: SPAMS optimizes an LASSO type objective iteratively. The number of iterations is set to be 1000 and the penalty parameter in front of the ℓ_1 norm is $\lambda = .1/\sqrt{N}$.
- DL-BCD: Our algorithm has an outer loop and an inner loop. The outer loop is set to be at most 3. The inner loop is allowed a maximum of 100 iterations. τ is either ∞ or 0.5.
- ER-SpUD: We use the default settings in the package developed by the authors of ER-SpUD.

First we introduce the simulation setting. We generate $n = 100p$ samples using a noisy linear model:

$$\boldsymbol{x}^{(i)} = \boldsymbol{D}^*\boldsymbol{\alpha}^{(i)} + \epsilon^{(i)}, \quad i = 1, \dots, n.$$

The reference dictionary $\boldsymbol{D}^*$, the reference coefficients $\boldsymbol{\alpha}^{(i)}$, and the noise $\epsilon^{(i)}$ are generated as follows.

- Generation of $\boldsymbol{D}^*$: First, we randomly generate a Gaussian matrix $X \in \mathbb{R}^{p\times p}$ where each entry X_{jk} is i.i.d. and $X_{jk} \sim \mathcal{N}(0, 1)$. We then scale columns of X to get columns of the reference dictionary $\boldsymbol{D}_j^* = X_j/\|X_j\|_2$ for $j = 1, \dots, p$.
- Generation of $\boldsymbol{\alpha}^{(i)}$: We generate the reference coefficient from sparse Gaussian distribution with sparsity s: $\boldsymbol{\alpha}^{(i)} \sim SG(s)$ for $i = 1, \dots, n$.
- Generation of $\epsilon^{(i)}$: We generate $\epsilon^{(i)}$ using a Gaussian distribution with mean zero. The variance of the distribution is set such that the signal-to–noise ratio is 100:

$$\frac{\mathbb{E}\|\boldsymbol{D}^*\boldsymbol{\alpha}^{(1)}\|_2}{\mathbb{E}\|\epsilon^{(1)}\|_2} = 10^2.$$

We choose the dimension p between 2 and 20 and sparsity s between 2 and p. For each (s, p)-pair, we repeat the experiment 100 times. The accuracy of an estimated dictionary $\hat{\boldsymbol{D}}$ is quantified using the normalized mean square error (NMSE):

$$\text{NMSE}(\hat{\boldsymbol{D}}, \boldsymbol{D}^*) = \min_{J\in\mathcal{J}} \frac{\|\hat{\boldsymbol{D}}J - \boldsymbol{D}^*\|_F^2}{\|\boldsymbol{D}^*\|_F^2},$$

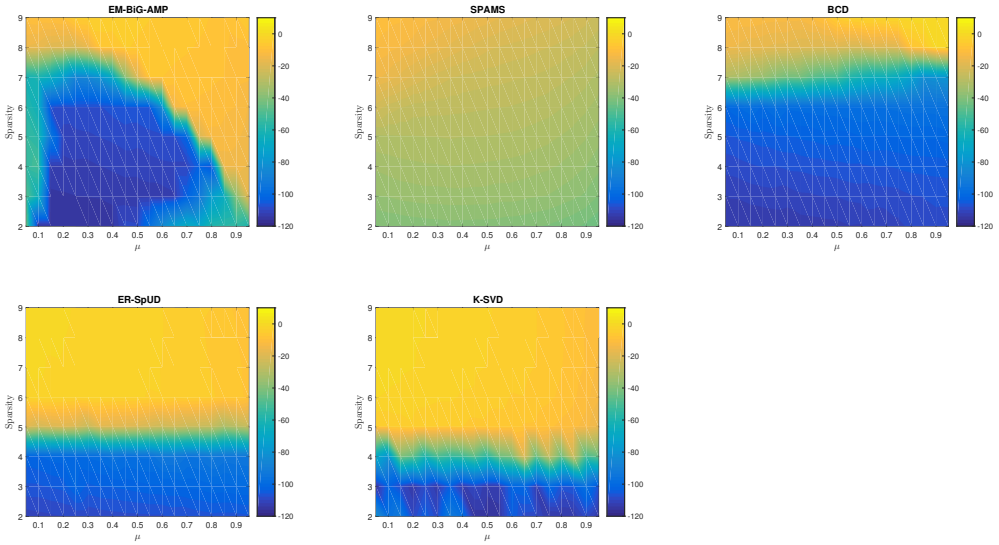


Figure 5.7: NMSE of recovered dictionaries using different algorithms. The reference dictionary is constant linearity matrix $(\mathbb{I} + \mu \mathbf{1})^{1/2}$. Coefficients are positive sparse Gaussian $\alpha \sim |\mathcal{S}(T)|$. Noise is standard Gaussian. SNR is 100dB.

where $\mathcal{J} = \{\Gamma \cdot \Lambda \mid \Gamma$ is a permutation matrix and Λ is a diagonal matrix whose diagonal elements are $\pm 1.\}$ is a set of matrices introduced to resolve the permutation and scale ambiguities. We say an algorithm has a successful recovery if the NMSE of $\hat{\boldsymbol{D}}$ is smaller than the threshold 0.01. We compare different algorithms in terms of their recovery rate, defined as the proportion of simulations that an algorithm has a successful recovery.

The algorithms being tested have several important parameters. For the purpose of comparison, we choose these parameters in a way such that they are consistent with other papers [68, 69]. The details of parameter settings can be found in Appendix D. We also added 40 dB noise.

Fig. 5.10 shows the recovery rate for a variety of choices of dimension p and sparsity s. For each algorithm, the blue region corresponds to (s, p) configurations under which an algorithm has high recovery rate, whereas yellow region indicates low recovery rate. Our results demonstrate that DL-BCD with $\tau = 0.5$ has the best recovery performance compared to other algorithms. We tried $\tau = 0.1, 0.5, 1, 2, 10$, and ∞ but with no further fine tuning. The algorithm EM-BiG-AMP has the second best performance.

We also compare the algorithms in terms of their computation cost. We record the average computation times for $p = 20$ and $s = 10$ (Fig. 5.11). It can be seen that the SPAMS package is the fastest. The speed of our DL-BCD is roughly the same as that of K-SVD. ER-SpUD is

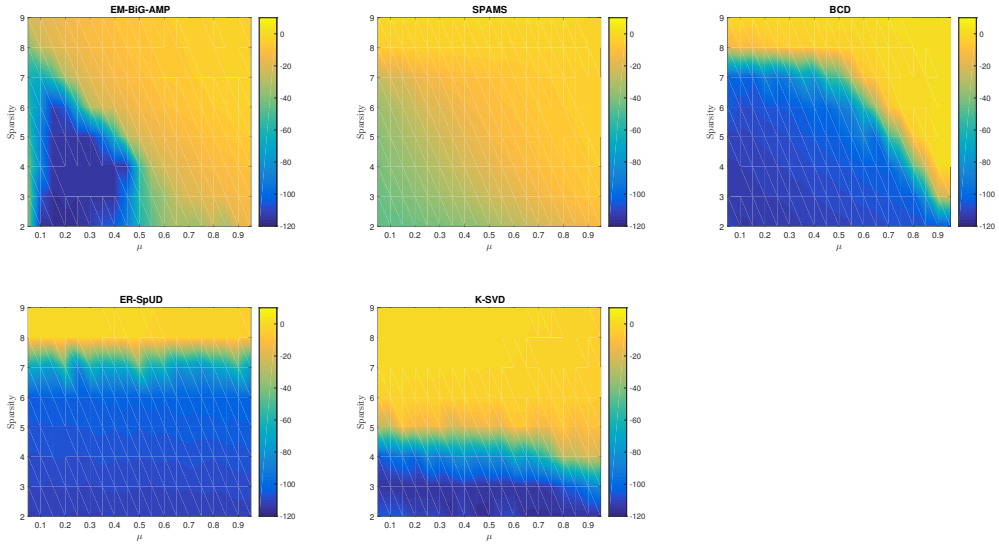


Figure 5.8: NMSE of recovered dictionaries using different algorithms. The reference dictionary is constant linearity matrix $(\mathbb{I} + \mu\mathbf{1})^{1/2}$. Coefficients are standard sparse Gaussian $\alpha \sim \mathcal{S}(T)$. Noise is standard Gaussian. SNR is 100dB.

the slowest among all the algorithms.

Drosophila embryonic gene expression images

To study organ formation of *Drosophila*, or fruit fly, developmental biologists use dye chemistry to visualize the gene expression of the fly embryos [33]. For early stage embryos, a study by [101] using nonnegative matrix factorization (NMF) reveals 21 principal patterns that correspond to different body parts and pre-organ regions. The data set contains 1640 gene expression images each of which is of dimension 32×16 pixels. A sample of original images can be found in Fig. 5.12. This dataset contains biologically interpretable image patterns that can be learned through Dictionary learning or NMF. In our following experiments, the dictionary size is chosen to be 21 in order to compare directly with [101].

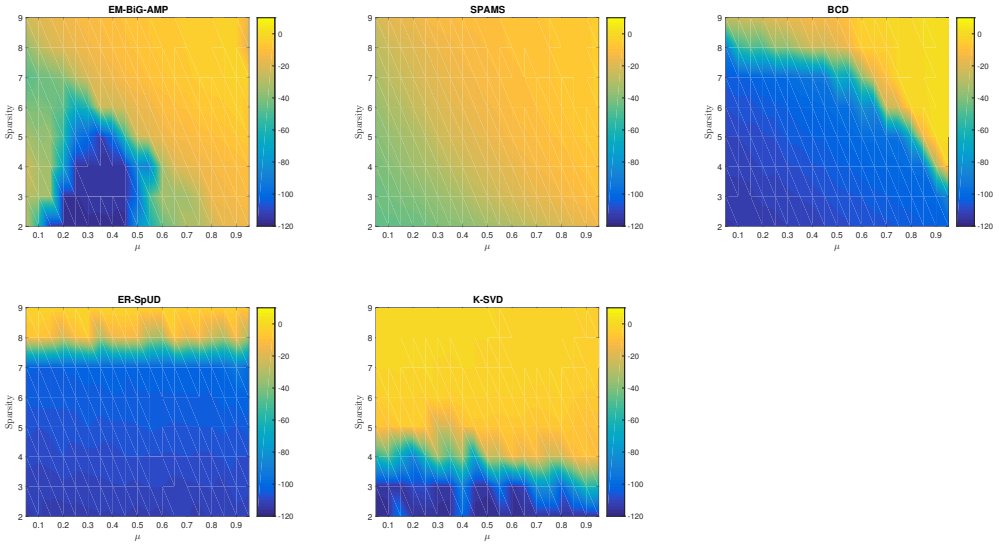


Figure 5.9: NMSE of recovered dictionaries using different algorithms. The reference dictionary is constant linearity matrix $(\mathbb{I} + \mu\mathbf{1})^{1/2}$. Coefficients are sparse Laplacian $\alpha \sim \mathrm{SL}(T)$. Noise is standard Gaussian. SNR is 100dB.

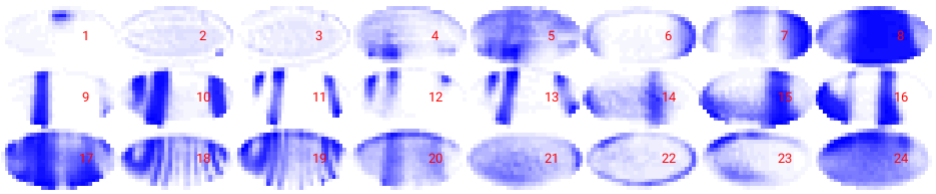


Figure 5.12: A sample of early stage *Drosophila* embryonic gene expression images. Each image corresponds to the spatial expression of one gene. There can be multiple replicates for the same gene in the data. The blue region is where the gene is expressed.

For this dataset, we are learning an under-complete dictionary. In order to apply DL-BCD, we first perform a PCA and select the first 21 principle components. The DL algorithms are then applied to the loading coefficients of the 21 PCs. Each algorithm is repeated 5 times and the best result is selected. NMF, SPAMS-DL, DL-BCD, K-SVD, EM-BiG-AMP, and ER-SpUD results are shown in Fig. 5.13. From the result, it can be seen that NMF gives the most interpretable result in the sense that all learned patterns are well localized in different

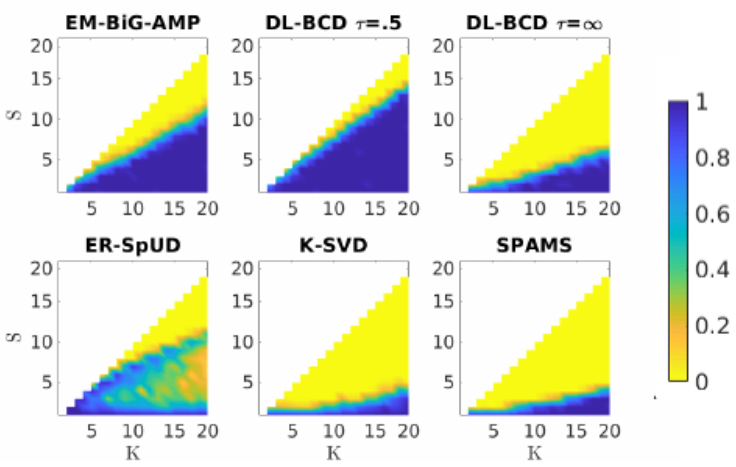


Figure 5.10: Recovery rate of different algorithms for $p = 2, \ldots, 20$ and $s = 2, \ldots, p$.

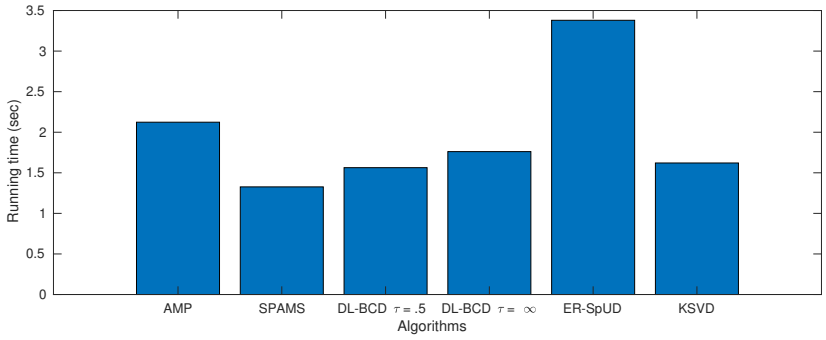
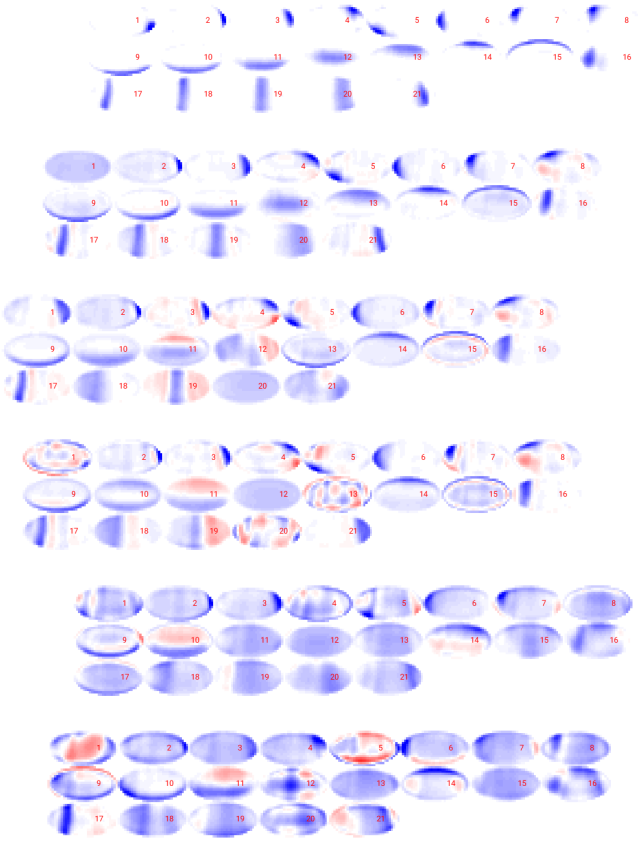


Figure 5.11: Average running time of different algorithms for $p = 20$ and $s = 10$.

geographic locations of the embryo. Then SPAMS-DL also gives dictionaries that are very similar to the NMF result. DL-BCD and EM-BiG-AMP both recover some of the atoms in the dictionary but there are certain patterns that are not recognizable. This shows that DL-BCD has reasonable performance even for noisy data. The ground truth dictionary is the NMF result.



(a): NMF

(b): SPAMS-DL

(c): DL-BCD ($\tau = \infty$)

(d): EM-BiG-AMP

(e): K-SVD

(e): ER-SpUD

Figure 5.13: Learned dictionaries using different DL algorithms. Positive values are painted blue while negative values are painted red.

Chapter 6

Conclusion and future work

6.1 Conclusion

In the first part of the thesis, we study the theoretical properties of ℓ_1-minimization dictionary learning under complete reference dictionary and noiseless signal assumptions. First, we derive a sufficient and almost necessary condition for local identifiability of ℓ_1-minimization. Our theorems not only extend previous local identifiability results to a much wider class of coefficient distributions, but also give an explicit bound on the region within which the objective value of the reference dictionary is minimal and characterize the sharpness of a local minimum. Second, we show that the reference dictionary is the unique sharp local minimum for ℓ_1-minimization. Based on our theoretical results, we design an algorithm to check the sharpness of a local minimum numerically. Finally, We propose the DL-BCD algorithm and demonstrate its competitive performance over other state-of-the-art algorithms in noiseless complete dictionary learning.

6.2 Future directions

Although we mainly focus on complete dictionaries, we believe that some of the results can be extended to the over-complete case. The challenge is that the representation of the optimization problem in the complete case (Formulation 4.1) will become much more complicated as the dictionary is no longer invertible. To deal with this issue, we note that some collections of the columns of the dictionary are invertible and as a result, the problem is now a double minimization $\min_{\boldsymbol{D}\in\mathcal{D}} \mathbb{E} \min_{\boldsymbol{D}'\in\mathbb{R}^{p\times p},\boldsymbol{D}'\subset\boldsymbol{D}} \|\boldsymbol{D}'^{-1}\boldsymbol{x}\|_1$. Techniques used in compressed sensing [18, 28] and prior works of overcomplete dictionary learning [30] can be useful in establishing the generalized results. Besides over-complete settings, it would also be interesting to generalize the result to the noisy case [31].

Although our results only apply to complete dictionaries, the insights gained from the analysis has potential to be generalized to the over-complete dictionary. Under appropriate

models, we conjecture that the possibly overcomplete reference dictionary could be the only sharp local minimum for the ℓ_1 minimization objective.

Part II

Feature importance and feature interaction recovery via RF

Chapter 7

Feature importance and feature interaction discovery: Introduction

Machine learning algorithms have been empirically proven to be extremely powerful in terms of prediction accuracy for various supervised learning tasks. However, prediction accuracy is not the only goal in many scientific and industrial problems. Information about how a model makes predictions is of paramount value [62]. For example, when a geneticist wants to understand a particular phenotype, e.g., hair color, having an black-box algorithm that *predicts* hair color from someone's genotype does not provide any new *biological insights*. Instead, the primary focus of interest is often the mechanism behind the data itself – what genes are important and what genes interact with each other for a particular phenotype? Thus, in these research areas, it is a pressing task to interpret the ML models and extract information beyond prediction.

Among many ML algorithms, tree ensembles including Random Forests (RF) [12] and gradient boosted decision trees [27] stand out as they enjoy both state-of-the-art prediction performance in a variety of practical problems and have relatively reliable interpretation algorithms [86, 56, 103, 53, 50]. To interpret a tree ensemble model, two questions are at the center:

- **Feature importance**: What features are important for the model's prediction?
- **Feature interaction**: How do features interact with one another to form the final prediction?

In the following section, we give a high-level overview of works in each of these directions.

RF feature importance

Understanding how a machine learning (ML) model makes predictions is important in many scientific and industrial problems. Appropriate interpretations can help increase the predictive performance of a model and provide new domain insights. While a line of study focuses

on interpreting any generic ML model [89, 74], there is a growing interest in developing specialized methods to understand specific models. In particular, interpreting Random Forests (RF) [12] and its variants [54, 88, 86, 87, 10, 44] has become an important area of research due to the wide ranging applications of RF in various scientific areas, such as genome-wide association studies (GWAS) [24], gene expression microarray [47, 75], and gene regulatory networks [38].

A key question in understanding RF is how to assign feature importance. That is, which features does a RF rely on for prediction? One of the most widely used feature importance measures for RF is mean decrease impurity (MDI) [13]. MDI computes the total reduction in loss or impurity contributed by all splits for a given feature. This method is computationally very efficient and has been widely used in a variety of applications [79, 38]. However, theoretical analysis of MDI has remained sparse in the literature [41]. Assuming there are an infinite number of samples, Louppe et al. [56] characterized MDI for totally randomized trees using mutual information between features and the response. They showed that noisy features, i.e., features independent of the outcome, have zero MDI importance. However, empirical studies have shown that MDI systematically assigns higher feature importance values to numerical features or categorical features with many categories [87]. In other words, high MDI values do not always correspond to the predictive associations between features and the outcome. We call this phenomenon MDI *feature selection bias*. Louppe [55] studied this issue and demonstrate via simulations that early stopping mechanisms (e.g., limited depth and larger leaf sizes) are effective to reduce the feature selection bias.

In addition to MDI [96, 57], some other feature importance measures have been studied in the literature and used in practice:

- Split count, namely, the number of times a feature is used to split [87], can be used as a feature importance measure. This method has been studied in [88, 10] and is available in XGBoost [19].

- Mean decrease in accuracy (MDA) measures a feature's importance by the reduction in the model's accuracy after randomly permuting the values of a feature. The motivation of MDA is that permuting an important feature would result in a large decrease in the accuracy while permuting an unimportant feature would have a negligible effect. Different permutation choices have been studied in [88, 39].

Recently, Lundberg et al. [57] show that for feature importance measures such as MDI and split counts, the importance of a feature does not always increase as the outcome becomes more dependent on that feature. To remedy this issue, they propose the tree SHAP feature importance, which focuses on giving consistent feature attributions to each sample. When individual feature importance is obtained, overall feature importance is straightforward to obtain by just averaging the individual feature importances across samples.

There is another line of work that focuses on modifying the tree construction procedure to obtain better feature importance measures. Hothorn et al. [36] introduced cforest in the R package `party` that grows classification trees based on a conditional inference framework.

Strobl et al. [87] showed that cforest suffers less from the feature selection bias. Sandri and Zuccolotto [79] proposed to create a set of uninformative pseudo-covariates to evaluate the bias in Gini importance. Nembrini et al. [64] gave a modified algorithm that is faster than the original method proposed by Sandri and Zuccolotto [79] with almost no overhead over the creation of the original RF and available in the R package `ranger`. In a very recent paper, Zhou and Hooker [103] proposed to evaluate the decrease in impurity at each node using out-of-bag samples. However, our implementation is different from that in [103] and MDI-oob enjoys higher computational efficiency.

Feature interaction discovery

In gene-interactions studies (also called epistasis), some papers [95, 102] empirically analyze the extraction of feature interactions from paths of ensembles of decision trees. Wan et al. [95] consider a boosting algorithm called MegaSNPHunter, where they interpret all groups of features that jointly appear on one of the decision paths as a candidate interaction. However, for genome-wide data this approach is computationally challenging even for mid-sized trees because the algorithm generates a massive list of candidate interactions. Moreover, MegaSNPHunter does not derive a ranking of those candidate interactions from the tree structure. Yoshida and Koike [102] propose to rank candidate interactions of genetic variants based on how often they appear together on decision paths in RF. However, this approach requires a brute-force search among all possible feature combinations, thus can only be applied to a relatively small number of features. Recently, iterative Random Forests (iRF) [10] is proposed to seek predictive, stable, and high-order feature interactions based on a similar idea as in Yoshida and Koike [102] that the set of interacting features often appear together on individual decision paths of a tree. However, iRF incorporates a soft dimension reduction step via iterative re-weighting of features in terms of their Gini importances, in order to stabilize individual decision paths in the trees. Using the random intersection trees (RIT) algorithm, iRF can extract stable interactions of arbitrary order in a computationally efficient way, even when the number of features is large. There is positive evidence that iRF extracts predictive, stable, and high-order interaction information from RF in genomics and other fields [10, 45].

While these works provide strong empirical evidence that interactions extracted from the ensemble of decision trees via RF are informative about underlying biological functional relationships, the theoretical foundation of tree-based methods remain unexamined. In the following chapters, we aim to provide theoretical understanding into RF based algorithms for both feature importance and feature interaction discovery. We study the problem of RF feature importance in Chapter 8. Then we introduce the LSS model and our main theoretical results on feature interaction discovery in Chapter 9. Inspired by our theoretical results on LSS model, we propose a novel ranking criterion called LSSrank and evaluate its empirical performance in Chapter 10. Finally, we conclude our work in Chapter 11.

Chapter 8

Debiased feature importance via out-of-bag samples

In this chapter, using the original definition of MDI, we analyze the non-asymptotic behavior of MDI and bridge the gap between the population case and the finite sample case. We find that under mild conditions, if the samples used for each tree are i.i.d, then the expected MDI feature importance of noisy features derived from any tree ensemble constructed on n samples with p features is upper bounded by $d_n \log(np)/m_n$, where m_n is the minimum leaf size and d_n is the maximum tree depth in the ensemble. In other words, deep trees with small leaves suffer more from feature selection bias. Our findings are particularly relevant for practical applications involving RFs, in which scenario deep trees are recommended [12] and used more often. To reduce the feature selection bias for RFs, especially when the trees are deep, we derive a new analytical expression for MDI and then use this new expression to propose a new feature importance measure that evaluates MDI using out-of-bag samples. We call this importance measure MDI-oob. For both regression and classification problems, we use simulated data and a genomic dataset to demonstrate that MDI-oob often achieves 5%–10% higher AUC scores compared to other feature importance measures used in several publicly available packages including `party` [16], `ranger` [99], and `scikit-learn` [70].

The rest of the chapter is organized as follows: we first provide a non-asymptotic analysis to quantify the bias in the MDI importance when noisy features are independent of relevant features in Section 8.1. In Section 8.2, we give a new characterization of MDI and propose a new MDI feature importance using out-of-bag samples, which we call MDI-oob. In Section 8.3, we compare our MDI-oob with other commonly used feature importance measures in terms of feature selection accuracy using the simulated data and a genomic ChIP dataset. We conclude our work and discuss possible future directions in Section 8.4.

8.1 Understanding the feature selection bias of MDI

In this section, we focus on understanding the finite sample properties of MDI importance and why it may have a significant bias in feature selection. We first briefly review the construction of RFs and introduce some important notations. Then, using the original definition of MDI, we give a tight upper bound to quantify the expected bias of MDI importance for a noisy feature. This upper bound is tight up to a $\log n$ factor where n is the number of i.i.d. samples.

Background and notations

Suppose that the data set $\mathcal{D}$ contains n i.i.d samples from a random vector $(X_1, \ldots, X_p, Y)$, where $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ are p input features and $Y \in \mathbb{R}$ is the response. The i^{th} sample is denoted by $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. We say that a feature X_k is a *noisy* feature if X_k and Y are independent, and a *relevant* feature otherwise. Note that this definition of noisy features has also been used in many previous papers such as [56, 83]. We denote $S \subset [p]$ as the set of indexes of relevant features. We are particularly interested in the case where the number of relevant features is small, namely, $|S|$ is much smaller than p. For any number $m \in \mathbb{N}$, $[m]$ denotes the set of integers $\{1, \ldots, m\}$. For any hyper-rectangle $R \subset \mathbb{R}^p$, let $\mathbf{1}(X \in R)$ be the indicator function taking value one when $X \in R$ and zero otherwise.

RFs are an ensemble of classification and regression trees, where each tree T defines a mapping from the feature space to the response. Trees are constructed independently of one another on a bootstrapped or subsampled data set $\mathcal{D}^{(T)}$ of the original data $\mathcal{D}$. Any node t in a tree T represents a subset (usually a hyper-rectangle) R_t of the feature space. A split of the node t is a pair (k, z) which divides the hyper-rectangle R_t into two hyper-rectangles $R_t \cap \mathbf{1}(X_k \leq z)$ and $R_t \cap \mathbf{1}(X_k > z)$, corresponding to the left child t^{left} and right child t^{right} of node t, respectively. For a node t in a tree T, $N_n(t) = |\{i \in \mathcal{D}^{(T)} : \mathbf{x}_i \in R_t\}|$ denotes the number of samples falling into R_t and

$$\mu_n(t) := \frac{1}{N_n(t)} \sum_{i:\mathbf{x}_i \in R_t} y_i \tag{8.1}$$

denotes their average response.

Each tree T is grown using a recursive procedure which proceeds in two steps for each node t. First, a subset $\mathcal{M} \subset [p]$ of features is chosen uniformly at random. Then the optimal split $v(t) \in \mathcal{M}, z(t) \in \mathbb{R}$ is determined by maximizing:

$$\Delta_{\mathcal{I}}(t) := \text{Impurity}(t) - \frac{N_n(t^{\text{left}})}{N_n(t)} \text{Impurity}(t^{\text{left}}) - \frac{N_n(t^{\text{right}})}{N_n(t)} \text{Impurity}(t^{\text{right}}) \tag{8.2}$$

for some impurity measure $\text{Impurity}(t)$. The procedure terminates at a node t if two children contain too few samples, i.e., $\min\{N_n(t^{\text{left}}), N_n(t^{\text{right}})\} \leq m_n$, or if all responses are identical. The threshold m_n is called the *minimum leaf size*. If a node t does not have any children, it is called a leaf node; otherwise, it is called an inner node. We define the set of inner nodes of

a tree T as $I(T)$. We say that T' is a sub-tree of T if T' can be obtained by pruning some nodes in T.

Some popular choices of the impurity measure Impurity(t) include variance, Gini index, or entropy. For simplicity, we focus on the variance of the responses, i.e.,

$$\text{Impurity}(t) = \frac{1}{N_n(t)} \sum_{i:\mathbf{x}_i \in R_t} (y_i - \mu_n(t))^2, \tag{8.3}$$

throughout the thesis unless stated otherwise. Later we show that this definition of impurity is equivalent to the Gini index of categorical variables with one hot encoding (see Remark in Section 8.2)

The Mean Decrease Impurity (MDI) feature importance of X_k, with respect to a single tree T (first proposed by Breiman et al. in [13]) and an ensemble of n_{tree} trees $T_1, \ldots, T_{n_{tree}}$, can be written as

$$\text{MDI}(k, T) = \sum_{t \in I(T), v(t)=k} \frac{N_n(t)}{n} \Delta_{\mathcal{I}}(t) \quad \text{and} \quad \text{MDI}(k) = \frac{1}{n_{tree}} \sum_{s=1}^{n_{tree}} MDI(k, T_s), \tag{8.4}$$

respectively. This expression is the best known formula for MDI and was analyzed in many papers such as Louppe et al. [56].

Finite sample bias of MDI importance for RF

Given the set S of relevant features and a tree T, we denote

$$G_0(T) = \sum_{k \notin S} \text{MDI}(k, T) \tag{8.5}$$

as the sum of MDI importance of all noisy features. Ideally, $G_0(T)$ should be close to zero with high probability, to ensure that no noisy features get selected when using MDI importance for feature selection. In fact, Louppe et al. [56] show that $G_0(T)$ is indeed zero almost surely if we grow totally randomized trees with infinite samples. However, $G_0(T)$ is typically non-negligible in real data, and finite sample properties of $G_0(T)$ are not well understood. In order to bridge this gap, we conduct a non-asymptotic analysis of the expected value of $G_0(T)$. Our main result characterizes how the expected value of $G_0(T)$ depends on m_n, the minimum leaf size of T, and p, the dimension of the feature space. We start with the following simple but important fact.

Proposition 8.1.1. *If T' is a sub-tree of T, then* $\text{MDI}(k, T') \leq \text{MDI}(k, T)$ *for any feature X_k.*

This fact naturally follows from the observation that by definition, $\Delta_{\mathcal{I}}(t) \geq 0$ for any node t. Since the impurity decrease at each node is guaranteed to be non-negative, $G_0(T)$ will never decrease as T grows deeper, in which case the minimum leaf size m_n will be smaller.

Indeed, if T is grown to purity ($m_n = 1$), and all features are noisy ($S = \emptyset$), then $G_0(T)$ would simply be equal to the sample variance of the responses in the data $\mathcal{D}^{(T)}$. How fast does $G_0(T)$ increase as the minimum leaf size m_n becomes smaller? To quantify the relation between $G_0(T)$ and m_n, we need a few mild conditions which we now describe. Let

$$y_i = \phi(\mathbf{x}_{i,S}) + \epsilon_i, i = 1, \ldots, n \tag{8.6}$$

for some unknown function $\phi : \mathbb{R}^{|S|} \to \mathbb{R}$, where ϵ_i are i.i.d zero-mean Gaussian noise. We make the following assumptions.

(A1) $X_k \sim \text{Unif}[0, 1]$ for all $k \in [p]$. In addition, the noisy features $\{X_k, k \in [p]\backslash S\}$ are mutually independent, and independent of all relevant features. Here S denotes the set of relevant features.

(A2) ϕ is bounded: $\sup_{\mathbf{x}\in[0,1]^{|S|}} |\phi(\mathbf{x})| \leq M$ for some $M > 0$.

The Assumptions (A1) and (A2) are weaker than the assumptions usually made when studying the statistical properties of RF. The marginal uniform distribution condition in (A1) is common in the RF literature [83], and can be easily satisfied by transforming the features via its inverse CDF. Since we are interested in characterizing the MDI of noisy features, we do not require the relevant features to be independent of each other. We do require that noisy features are independent of relevant features, which is a limitation of Theorem 8.1.1 below. Correlated features are commonly encountered in practice and difficult for any feature selection method.

We now state our first main result which provides a non-asymptotic upper and lower bound for the expected value of the maximum of $G_0(T)$ over all tree T with minimum leaf size m_n.

Theorem 8.1.1. *Let $\mathcal{T}_n(m_n)$ denote the set of decision trees whose minimum leaf size is lower bounded by m_n, and $\mathcal{T}_n(m_n, d_n) \subset \mathcal{T}_n(m_n)$ denote the subset of $\mathcal{T}_n(m_n)$ whose depth is upper bounded by d_n. Under Assumptions (A1) and (A2), there exists a positive constant C such that,*

$$\mathbb{E}_{X,\epsilon} \sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T) \leq C\frac{d_n \log(np)}{m_n}. \tag{8.7}$$

In addition, when $f = 0$ and $m_n \geq 36 \log p + 18 \log n$,

$$\mathbb{E}_{X,\epsilon} \sup_{T\in\mathcal{T}_n(m_n)} G_0(T) \geq \frac{\log p}{Cm_n}. \tag{8.8}$$

We give the proof in the Appendix. To the best of our knowledge, Theorem 8.1.1 is the first non-asymptotic result on the expected MDI importance of tree ensembles. In particular, the upper bound can be directly applied to *any* tree ensembles with a minimum leaf size m_n and a maximum tree depth d_n, including Breiman's original RF procedure, if subsampling is used instead of bootstrapping.

Proof Sketch. Every node t in a tree $T \in \mathcal{T}_n(m_n, d_n)$ corresponds to an axis-aligned hyper-rectangle in $[0, 1]^p$ which contains at least m_n samples and is formed by splitting on at

most d_n dimensions consecutively. Therefore, bounding the supremum of impurity reduction for any potential node in $\mathcal{T}_n(m_n, d_n)$ boils down to controlling the complexity of all such hyper-rectangles. Two hyper-rectangles are considered equivalent if they contain the same subset of samples, since the impurity reductions of these two hyper-rectangles are always the same. Up to this equivalence, it can be proved that the number of unique hyper-rectangles of interest is upper bounded by $(np)^{d_n}$, which corresponds to the $d_n log(np)$ term in the upper bound. The final result is obtained via union bound. □

In the upper bound, each node t is obtained by splitting on at most d_n features. In practice, d_n is typically at most of order $\log n$. Indeed, if the decision tree is a balanced binary tree, then $d_n \leq \log_2 n$. Therefore, for balanced trees, the upper bound can be written as

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq C \frac{d_n \log(np)}{m_n} \leq C \frac{(\log n)^2 + \log n \log p}{m_n}, \tag{8.9}$$

and the theorem shows that the sum of MDI importance of noisy features is of order $\frac{\log p}{m_n}$, i.e.,

$$\sup_{\phi: \|\phi\|_\infty \leq M} \mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n)} G_0(T) \sim \frac{\log p}{m_n}, \tag{8.10}$$

up to a $\log n$ term correction, which is typically small in the high dimensional $p \gg n$ setting. If all features X_j are categorical with a bounded number of categories, then the upper bound can be improved to

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq C \frac{d_n \log p}{m_n}, \tag{8.11}$$

which shows that the MDI importance of noisy features can be better controlled if the noisy features are categorical rather than numerical. That is consistent with the previous empirical studies because the number of candidate split points for a numerical feature is larger than that for a categorical feature.

Theorem 8.1.1 shows that the supremum of MDI importance of noisy features over all trees with minimum leaf size m_n is, in expectation, roughly inversely proportional to m_n. Fig. 8.5, we show that the inversely proportional relationship is consistent with the empirical $G_0(T)$ on a simulated dataset described in the first simulation study in Section 8.3. Therefore, to control the finite sample bias of MDI importance, one should either grow shallow trees, or use only the shallow nodes in a deep tree when computing the feature importance. In fact, since $G_0(T)$ depends on the dimension p only through a log factor $\log p$, we expect $G_0(T)$ to be very small even in a high-dimensional setting if m_n is larger than, say, $\sqrt{n}$. For a balanced binary tree grown to purity with depth $d_n = \log_2 n$, this corresponds to computing MDI only from the first $d_n/2 = (\log_2 n)/2$ levels of the tree, as the node size on the dth level of a balanced tree is $n/2^d$.

Fact 8.1.1 implies that the MDI importance of relevant features might also decrease as m_n increases, but we will show in simulation studies that they will decrease at a much slower rate, especially when the underlying model is sparse.

8.2 MDI using out-of-bag samples (MDI-oob)

As shown in the previous section, for balanced trees, the sum of MDI feature importance of all noisy features is of order $\frac{\log(p)}{m_n}$ if we ignore the $\log(n)$ terms. This means that the MDI feature selection bias becomes severe for trees with smaller leaf size m_n, which usually corresponds to a deeper tree. Fortunately, this bias can be corrected by evaluating MDI using out-of-bag samples. In this section, we first introduce a new analytical expression of MDI as the motivation of our new method, then we propose the MDI-oob as a new feature importance measure. For simplicity, in this section, we only focus on one tree T. However, all the results are directly applicable to the forest case.

A new characterization of MDI

Recall that the original definition of the MDI importance of any feature k is provided in Equation (8.4), that is, the sum of impurity decreases among all the inner nodes t such that $v(t) = k$. Although we can use this definition to analyze the feature selection bias of MDI in Theorem 8.1.1, this expression (8.4) gives us few intuitions on how we can get a new feature importance measure that reduces the MDI bias. Next, we derive a novel analytical expression of MDI, which shows that the MDI of any feature k can be viewed as the sample covariance between the response y_i and the function $f_{T,k}(\mathbf{x}_i)$ defined in Proposition 8.2.1. This new expression inspires us to propose a new MDI feature importance measure by using the out-of-bag samples.

Proposition 8.2.1. *Define the function $f_{T,k}(\cdot)$ to be*

$$f_{T,k}(X) = \sum_{t \in I(T): v(t)=k} \Big(\mu_n(t^{\text{left}}) \mathbf{1}(X \in R_{t^{\text{left}}}) + \mu_n(t^{\text{right}}) \mathbf{1}(X \in R_{t^{\text{right}}}) - \mu_n(t) \mathbf{1}(X \in R_t) \Big).$$

Then the MDI of the feature k in a tree T can be written as:

$$\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i) \cdot y_i, \tag{8.12}$$

Proof of Proposition 8.2.1. For simplicity, here we only present the proof for a single tree T. The case of multiple trees is straightforward. Recall that t^{left} and t^{right} are the left and right children of the node t. Based on (8.4), MDI at the node t is

$$\begin{aligned} \frac{N_n(t)}{|\mathcal{D}^{(T)}|} \Delta_{\mathcal{I}}(t) = \frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} & [y_i - \mu_n(t)]^2 \mathbf{1}(\mathbf{x}_i \in R_t) \\ & - [y_i - \mu_n(t^{\text{left}})]^2 \mathbf{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) - [y_i - \mu_n(t^{\text{right}})]^2 \mathbf{1}(\mathbf{x}_i \in R_{t^{\text{right}}}). \end{aligned} \tag{8.13}$$

Because $\mathbf{1}(\mathbf{x}_i \in R_t) = \mathbf{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) + \mathbf{1}(\mathbf{x}_i \in R_{t^{\text{left}}})$, the above term becomes

$$\begin{aligned}
&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}\left((y_i-\mu_n(t))^2-(y_i-\mu_n(t^{\text{left}}))^2\right)\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{left}}})\\
&\qquad+\left((y_i-\mu_n(t))^2-(y_i-\mu_n(t^{\text{right}}))^2\right)\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{right}}})\\
=&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))(2y_i-\mu_n(t)-\mu_n(t^{\text{left}}))\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{left}}})\\
&\qquad+(\mu_n(t^{\text{right}})-\mu_n(t))(2y_i-\mu_n(t)-\mu_n(t^{\text{right}}))\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{right}}}). \qquad (8.14)
\end{aligned}$$

Since $\sum_{i\in\mathcal{D}^{(T)}} y_i\mathbf{1}(\mathbf{x}_i \in t^{\text{left}}) = N_n(t^{\text{left}})\mu_n(t^{\text{left}})$, we know $\sum_{i\in\mathcal{D}^{(T)}}(y_i-\mu_n(t^{\text{left}}))\mathbf{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) = 0$. Similar equations hold for the right child t^{right}, too. Then (8.14) reduces to

$$\begin{aligned}
&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))(y_i-\mu_n(t))\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{left}}}) \qquad (8.15)\\
&\qquad+(\mu_n(t^{\text{right}})-\mu_n(t))(y_i-\mu_n(t))\mathbf{1}(\mathbf{x}_i\in R_{t^{\text{right}}}) \qquad (8.16)
\end{aligned}$$

Because of the definitions of $\mu_n(t^{\text{left}})$, $\mu_n(t^{\text{right}})$, and $\mu_n(t)$, we know

$$N_n(t^{\text{left}})\mu_n(t^{\text{left}}) + N_n(t^{\text{right}})\mu_n(t^{\text{right}}) = N_n(t)\mu_n(t). \qquad (8.17)$$

That implies $\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))\mathbf{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) + (\mu_n(t^{\text{right}})-\mu_n(t))\mathbf{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) = 0$. Using this equation, (8.16) can be written as

$$\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{left}}})+(\mu_n(t^{\text{right}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{right}}}). \qquad (8.18)$$

In summary, we have shown that:

$$\begin{aligned}
&\frac{N_n(t)}{|\mathcal{D}^{(T)}|}\Delta_{\mathcal{I}}(t)\\
=&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{left}}})+(\mu_n(t^{\text{right}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{right}}}).
\end{aligned}$$

Since the MDI of the feature k is the sum of $\frac{N_n(t)}{|\mathcal{D}^{(T)}|}\Delta_{\mathcal{I}}(t)$ across all inner nodes such that $v(t) = k$, we have

$$\begin{aligned}
&\sum_{t\in I(T)}\frac{N_n(t)}{|\mathcal{D}^{(T)}|}\Delta_{\mathcal{I}}(t)\mathbf{1}(v(t)=k)\\
=&\sum_{t\in I(T):v(t)=k}\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}(\mu_n(t^{\text{left}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{left}}})+(\mu_n(t^{\text{right}})-\mu_n(t))y_i\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{right}}})\\
=&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}\Big[\sum_{t\in I(T):v(t)=k}(\mu_n(t^{\text{left}})-\mu_n(t))\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{left}}})+(\mu_n(t^{\text{right}})-\mu_n(t))\mathbf{1}(\mathbf{x}_i\in\mathbb{R}_{t^{\text{right}}})\Big]y_i\\
=&\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}}f_{T,k}(\mathbf{x}_i)y_i.
\end{aligned}$$

That completes the proof. □

Although we have not seen this analytical expression in the prior works, we found that the functions $f_{T,k}(\cdot)$ have been studied from a quite different perspective. Those functions were first proposed in Saabas [78] to interpret the RF predictions for each individual sample. According to this paper, $f_{T,k}$ can be viewed as the "contribution" made by the feature k in the tree T. For any tree, those functions $f_{T,k}$ can be easily computed using the python package *treeinterpreter*.

It can be shown that $\sum_{i\in\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i) = 0$. That implies $\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i)\cdot y_i$ is essentially the sample covariance between $f_{T,k}(\mathbf{x}_i)$ and y_i on the bootstrapped dataset $\mathcal{D}^{(T)}$. This indicates a potential drawback of MDI: RFs use the training data $\mathcal{D}^{(T)}$ to construct the functions $f_{T,k}(\cdot)$, then MDI uses the same data to evaluate the covariance between y_i and $f_{T,k}(\mathbf{x}_i)$ in Equation (8.12).

Remark: So far we have only considered regression trees, and have defined the impurity at a node t using the sample variance of responses. For classification trees, one may use Gini index as the measure of impurity. We point out that these two definitions of impurity are actually equivalent when we use a one-hot vector to represent the categorical response. Therefore, our results are directly applicable to the classification case. Suppose that Y is a categorical variable which can take D values $c_1, c_2, \ldots, c_D$. Let $p_d = \mathbb{P}(Y = c_d)$. Then the Gini index of Y is $\text{Gini}(Y) = \sum_{d=1}^{D} p_d(1-p_d)$. We define the one-hot encoding of Y as a D-dimensional vector $\tilde{Y} = (\mathbf{1}(Y = c_1), \ldots, \mathbf{1}(Y = c_D))$. Then

$$\text{Var}(\tilde{Y}) = \|\tilde{Y} - \mathbb{E}\tilde{Y}\|_2^2 = \sum_{d=1}^{D}(\mathbb{E}\tilde{Y}_i^2 - (\mathbb{E}\tilde{Y}_i)^2) = \sum_{d=1}^{D}(\mathbb{E}\tilde{Y}_i - (\mathbb{E}\tilde{Y}_i)^2) = \sum_{d=1}^{D} p_d(1-p_d) = \text{Gini}(Y), \tag{8.19}$$

thereby showing that Gini index and variance are equivalent.

Evaluating MDI using out-of-bag samples

Proposition 8.2.1 suggests that we can calculate the covariance between y_i and $f_{T,k}(\mathbf{x}_i)$ in Equation (8.12) using the out-of-bag samples $\mathcal{D}\backslash\mathcal{D}^{(T)}$:

$$\text{MDI-oob of feature } k = \frac{1}{|\mathcal{D}\backslash\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}\backslash\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i)\cdot y_i. \tag{8.20}$$

In other words, for each tree, we calculate the $f_{T,k}(\mathbf{x}_i)$ for all the OOB samples $\mathbf{x}_i$ and then compute MDI-oob using (8.20). Although out-of-bag samples have been used for other feature importance measures such as MDA, to the best of the authors' knowledge, there are few results that use the out-of-bag samples to evaluate MDI feature importance. A naive way of using the out-of-bag samples to evaluate MDI is to directly compute the impurity decrease at each inner-node of a tree using OOB samples. However, this approach is not desirable since

the impurity decrease at each node is still always positive unless the responses of all the OOB samples falling into a node are constant. In this case, an argument similar to the proof of Theorem 1 can show that the bias of directly computing impurity using OOB samples could still be large for deep trees. The idea of MDI-oob depends heavily on the new analytical MDI expression. Without the new expression, it is not clear how one can use out-of-bag samples to get a better estimate of MDI. One highlight of the MDI-oob is its low computation cost. The time complexity of evaluating MDI-oob for RFs is roughly the same as computing the RF predictions for $|\mathcal{D}\backslash\mathcal{D}^{(T)}|$ number of samples.

8.3 Simulation experiments

Simulated study on the effect of minimum leaf size and the tree depth

In this simulation, we investigate the empirical relationship between MDI importance and the minimum leaf size. To mimic the major experiment setting in the paper [87], we generate the data as follows. We sample $n = 200$ observations, each containing 5 features. The first feature is generated from standard Gaussian distribution. The second feature is generated from a Bernoulli distribution with $p = 0.5$. The third/fourth/fifth features have 4/10/20 categories respectively with equal probability of taking any states. The response label y is generated from a Bernoulli distribution such that $P(y_i = 1) = (1 + x_{i2})/3$. While keeping the number of trees to be 300, we vary the minimum leaf size of RF from 1 to 50 and record the MDI of every feature. The results are shown in Fig. 8.1. We can see from this figure that the MDI of noisy features, namely X1, X3, X4 and X5, drops significantly when the minimum leaf size increases from 1 to 50. This observation supports our theoretical result in Theorem 8.1.1. Besides the minimum leaf size, we also investigate the relationship between MDI and the tree depth. As tree depth increases, the minimum leaf size generally decreases exponentially. Therefore, we expect the MDI of noisy features to become larger for increasing tree depth. We vary the maximum depth from 1 to 20 and record the MDI of every feature. The results shown in Fig. 8.2 are consistent with our expectation. MDI importance of noisy features increase when the tree depth increases from 1 to 20. Fig. 8.3 shows the MDI-oob measure and it indeed reduces the bias of MDI in this simulation.

Noisy feature identification using the simulated data

In this experiment, we evaluate different feature importance measures in terms of their abilities to identify noisy features in a simulated data set. We compare our method with the following methods: MDA, cforest in the R package `party`, SHAP[57], default feature importance (MDI) in `scikit-learn`, the impurity corrected Gini importance in the R package `ranger`, UFI in [103], and naive-oob, which refers to the naive method that evaluates impurity decrease at each node using out-of-bag samples directly. To evaluate feature importance measures, we generate the following simulated data. Inspired by the experiment settings in Strobl et al. [87], our setting involves discrete features with different number of distinct values, which poses

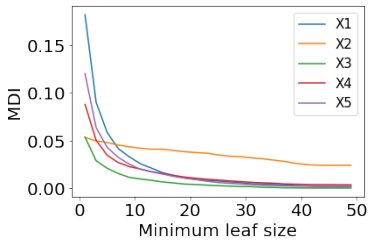


Figure 8.1: MDI against min leaf size.

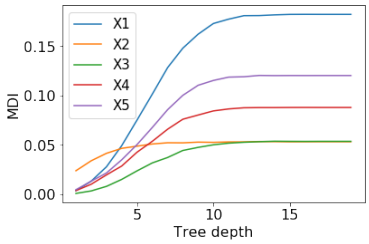


Figure 8.2: MDI against tree depth.

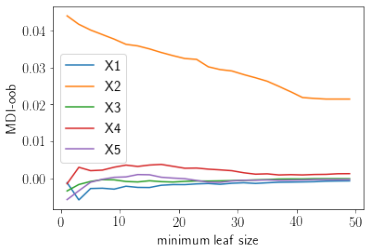


Figure 8.3: MDI-oob against min leaf size.

a critical challenge for MDI. The data has 1000 samples with 50 features. All features are discrete, with the j^{th} feature containing $j+1$ distinct values $0, 1, \ldots, j$. We randomly select a set S of 5 features from the first ten as relevant features. The remaining features are noisy features. Choosing active features with fewer categories represents the most challenging case for MDI. All samples are i.i.d. and all features are independent. We generate the outcomes using the following rules:

- Classification: $P(Y = 1|X) = \text{Logistic}(\frac{2}{5}\sum_{j\in S} X_j/j - 1)$.
- Regression: $Y = \frac{1}{5}\sum_{j\in S} X_j/j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 100 \cdot \text{Var}(\frac{1}{5}\sum_{j\in S} X_j/j))$.

Treating the noisy features as label 0 and the relevant features as label 1, we can evaluate a feature importance measure in terms of its area under the receiver operating characteristic curve (AUC). Note that when a feature importance measure gives low importance to relevant features, its AUC score measure can be smaller than 0.5 or even 0. We grow 100 trees with the minimum leaf size set to either 100 (shallow tree case) or 1 (deep tree case). The number of candidate features m_{try} is set to be 10. We repeat the whole process 40 times and report the average AUC scores for each method in Table 8.1. The boxplots For this simulated setting, MDI-oob achieves the best AUC score under all cases.

Noisy feature identification using a genomic ChIP dataset

To evaluate our method MDI-oob in a more realistic setting, we consider a ChIP-chip and ChIP-seq dataset measuring the enrichment of 80 biomolecules at 3912 regions of the *Drosophila* genome [17, 58]. These data have previously been used in conjunction with RF-based methods, namely Iterative Random Forest (iRF) [10], to predict functional labels associated with genomic regions. They provide a realistic representation of many issues encountered in practice, such as heterogeneity and dependencies among features, which make it especially challenging for feature selection problems. To evaluate feature selection in the ChIP data, we scale each feature X_j to be between 0 and 1. Second, we randomly select a

set S of 5 features as relevant features and include the rest as noisy features. We randomly permute values of any noisy features to break their dependencies with relevant features. By this means, we avoid the cases where RFs "think" some features are important not because they themselves are important but because they are highly correlated with other relevant features. Then we generate responses using the following rules:

- Classification:$P(Y = 1|X) = \text{Logistic}(\frac{2}{5}\sum_{j\in S} X_j - 1)$.
- Regression: $Y = \frac{1}{5}\sum_{j\in S} X_j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 100 \cdot \text{Var}(\frac{1}{5}\sum_{j\in S} X_j))$.

All the other settings remain the same as the previous simulations. We report the average AUC scores for each method in Table 8.1. The standard errors and the beeswarm plots of all the methods are included in the 8.4. Naive-oob, namely, the method that directly computes MDI using the out-of-bag samples is hardly any better than the original gini importance. MDI-oob or UFI usually achieves the best AUC score in three out of four cases, except for shallow regression trees, when all methods appear to be equally good with AUC scores close to 1. Although UFI and MDI-oob use out-of-bag samples in different ways, their results are generally comparable. We also note that increasing the minimum leaf size consistently improves the AUC scores of all methods.

Another observation is that MDA behaves poorly in some simulations despite its use of a validation set. This could be due to the low signal-to-noise ratio in the simulation setting. After we train the RF model on the training set, we evaluated the model's accuracy on a test set. It turns out that the accuracy of the model is quite low. In that case, MDA struggles because the accuracy difference between permuting a relevant feature and permuting a noisy feature is small. We observe that the MDA gets better when we increase the signal-to-noise ratio.

The computation time of different methods is hard to compare due to a few factors. Because the packages including `scikit-learn` and `ranger` compute feature importance when constructing the tree, it is hard to disentangle the time taken to construct the trees and the time taken to get the feature importance. Furthermore, different packages are implemented in different programming languages so it is not clear if the time difference is because of the algorithm or because of the language. We implement MDI-oob in Python and for our first simulated classification setting, MDI-oob takes $\sim$ 3.8 seconds for each run. To compare, `scikit-learn` which uses Cython (A C extension for Python) takes $\sim$ 1.4 seconds to construct the RFs for each run. Thus, MDI-oob runs in a reasonable time frame and we expect it to be faster if it is implemented in C or C++.

8.4 Discussion and future directions

Mean Decrease Impurity (MDI) is widely used to assess feature importance and its bias in feature selection is well known. Based on the original definition of MDI, we show that its expected bias is upper bounded by an expression that is inversely proportional to the

Table 8.1: Average AUC scores for noisy feature identification

	Deep tree (min leaf size = 1)				Shallow tree(min leaf size = 100)			
	Simulated		ChIP		Simulated		ChIP	
	C	R	C	R	C	R	C	R
MDI-oob	**0.76**	0.52	0.87	0.98	**0.75**	**0.58**	**0.94**	0.98
UFI	0.72	**0.54**	**0.88**	**0.99**	**0.75**	0.56	**0.94**	0.98
naive-oob	0.18	0.10	0.67	0.71	0.60	0.39	0.89	0.97
SHAP	0.55	0.33	0.82	0.96	0.68	0.46	0.91	0.97
ranger	0.56	0.50	0.73	0.97	0.55	0.49	0.76	**0.99**
MDA	0.49	0.51	0.54	0.97	0.50	**0.58**	0.50	**0.99**
cforest	0.65	0.50	0.79	0.93	0.70	0.49	0.90	0.98
MDI	0.12	0.09	0.60	0.71	0.63	0.40	0.88	0.97

"C" stands for classification, "R" stands for regression. The column maximum is bolded.

minimum leaf size under mild conditions, which means deep trees generally have a higher feature selection bias than shallow trees. To reduce the bias, we derive a new analytical expression for MDI and use the new expression to obtain MDI-oob. For the simulated data and a genomic ChIP dataset, MDI-oob has exhibited the state-of-the-art feature selection performance in terms of AUC scores.

Comparison to SHAP. SHAP originates from game theory and offers a novel perspective to analyze the existing methods. While it is desirable to have 'consistency, missingness and local accuracy', our analysis indicates that there are other theoretical properties that are also worth taking into account. As shown in our simulation, the feature selection bias of SHAP increases with the depth of the tree, and we believe SHAP can also use OOB samples to improve feature selection performance.

Relationship to honest estimation. Honest estimation is an important technique built on the core notion of sample splitting. It has been successfully used in causal inference and other areas to mitigate the concern of over-fitting in complex learners due to usage of same data in different stages of training. The proposed algorithm MDI-oob has important connections with "honest sampling" or "honest estimation". For example, in Breiman's 1984 book [13], he proposed to use a separate validation set for pruning and uncertainty estimation. In [94], each within-leaf prediction is estimated using a different sub-sample (such as OOB sample) than the one used to decide split points. Theoretical results of these papers and Proposition 8.2.1 convey the same message, that finite sample bias is caused by using the same data for growing trees and for estimation, and the bias can be reduced if we leverage OOB data. We believe the theoretical contributions of those papers can also help us analyze the statistical properties (such as variance) of the MDI-oob.

Future directions. Although the MDI-oob shows promising results for selecting relevant features, it also raises many interesting questions to be considered in the future. First of all, how can MDI-oob be extended to better accommodate correlated features? Going beyond

feature selection, can importance measures also rank the relevant features in a reasonable order? Finally, can we use the new analytical expression of MDI to give a tighter theoretical bound for MDI's feature selection bias? We are exploring these interesting questions in our ongoing work.

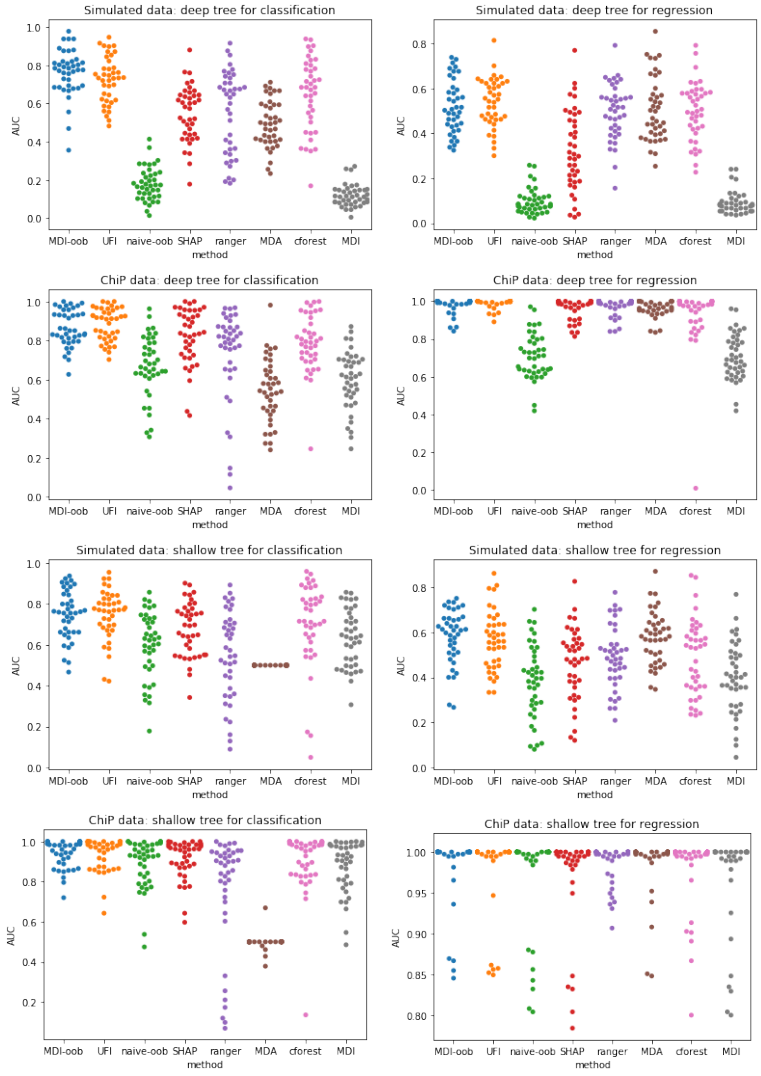


Figure 8.4: Beeswarm plots for different simulation settings described in Section 8.3. The left figures show AUC scores for different feature importance measures for classification problems. The right figure show AUC scores for different feature importance measures for regression problems. Both simulated data and ChiP data are considered for shallow and deep trees. In general, MDI-oob has highest average AUC across those settings.

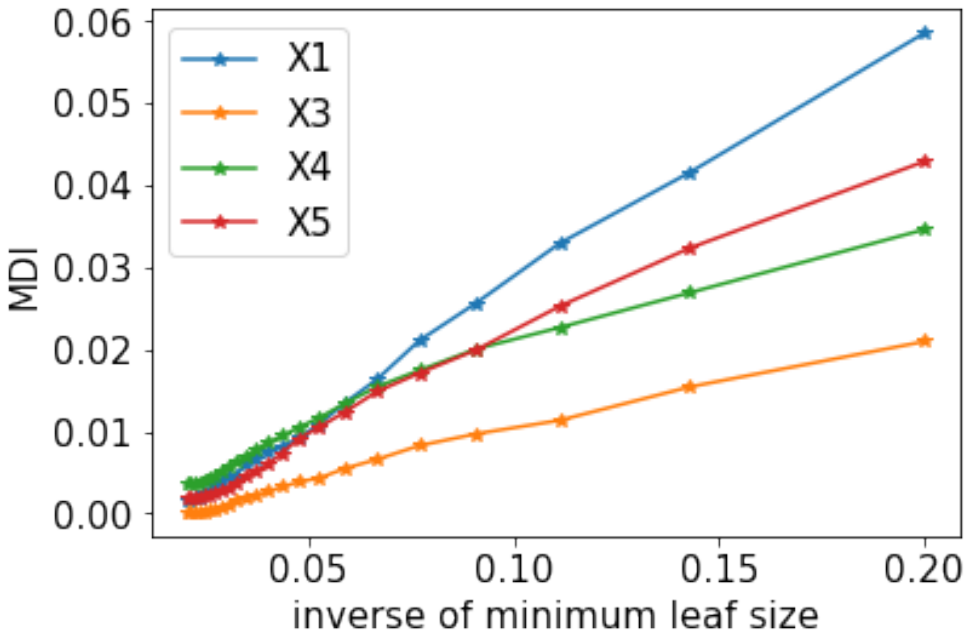


Figure 8.5: MDI against inverse min leaf size. This is coherent with our theoretical analysis as MDI is proportional to the inverse of minimum leaf size.

Chapter 9

Provable high-order interaction recovery

Theoretical results are relevant to practice if they inform and guide practice. They can not possibly be relevant if the theoretical data generation models do not capture reality in any domain area. One of the most common assumptions made in previous theoretical analyses of RF is a Lipschitz condition on the underlying mean regression function, see e.g., [11, 94], which, to our best knowledge, has not been supported by empirical evidence in high-dimensions. Moreover, many biological processes show thresholding or discontinuous interacting behavior among biomolecules [98, 35], which strongly violates the Lipschitz assumption. Motivated by this scientific observation, we propose the Local-Spiky Sparse (LSS) model: an additive Boolean interaction model with bounded noise.[1] LSS is inspired by genomics data problems where RF have shown impressive prediction performance [40, 20, 92]. Since Boolean functions are not continuous, the LSS model does not satisfy the Lipschitz assumption. However, it not only makes the theoretical model more relevant for biologists, but also matches the decision tree structure in RF analytically. Furthermore, the LSS model is able to capture spatial inhomogeneity in addition to discontinuous thresholding behavior. We believe it is suitable and useful as a new benchmark model under which to evaluate theoretically (and computationally) interaction discovery performance of ML algorithms including RF. For any tree ensembles, we define a new quantity called depth-weighted prevalence (DWP) on decision paths of a set of features. We show that DWP of RF has a universal upper bound that depends only on the size of the set. Furthermore, the upper bound is attained with high probability as the sample size increases if and only if the signed features represent an union of interactions in the LSS model. That implies that one can use RF to consistently recover interaction components in the LSS model regardless of the model coefficients. Our theoretical results show that feature subsampling of RF is essential to recover interactions. When too few features are sampled at each node, the prevalence of true interactions can be too small; When too many features are sampled, the prevalence of false interactions can be too high. More precisely, our results indicate that one needs to sample a constant fraction of features in order to learn higher-order interactions from tree paths. This also suggests that extremely

[1] The LSS model was first considered by authors of [10] (including one of us) and already used to evaluate the performance of iRF/siRF in [45].

randomized trees may not be ideal for interaction discovery as features used along the paths are purely random.

9.1 Local-Spiky Sparse (LSS) model: Boolean interactions

For an integer $N \in \mathcal{N}$, let $[N] := \{1, 2, \ldots, N\}$. For a set S of finite elements, let $|S|$ denote its cardinality or the number of elements in S. For any event A, $\mathbf{1}(A)$ denotes the indicator function of A. We consider the following data generating model.

LSS model. *Assume a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ of n samples. $\mathbf{x}_i = (x_{i1}, \ldots, x_{in}) \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ are i.i.d. samples from a distribution $P(X, Y)$ such that for some fixed constants $C_\beta > 0, C_\gamma \in (0, 0.5)$, and fixed integer $s \in \mathcal{N}$,*

1. *(Uniformity) X is uniformly distributed on $[0, 1]^p$;*
2. *(Bounded-response) Y is bounded, i.e. $|Y| < 1$;*
3. *(LSS-expectation) the regression function is*

$$E(Y|X) = \beta_0 + \sum_{j=1}^{J} \beta_j \prod_{k \in S_j} \mathbf{1}(X_k \gtrless \gamma_k) \tag{9.1}$$

where $S_1, \ldots, S_J \subset [p]$ are disjoint sets called basic interactions*, i.e.,*

$$S_{j_1} \cap S_{j_2} = \emptyset \text{ for all } j_1 \neq j_2;$$

$\cup_{j=1}^{J} S_j$ have at most s features, i.e.

$$\sum_{j=1}^{J} |S_j| \leq s;$$

coefficients β_j are bounded from below or $\min_{j=1}^{J} |\beta_j| > C_\beta > 0$ and thresholds γ_j are bounded away from 0 and 1, i.e., $\gamma_j \in (C_\gamma, 1 - C_\gamma)$, for $j = 1, \ldots, J$ and some $C_\gamma > 0$.

Here, $\gtrless$ in (9.1) means either $\leq$ or $\geq$, potentially different for every k. This inequality defines the sign of a feature in the interaction which will be defined more precisely in Definition 9.1.1. We associate $\leq$ in (9.1) with a negative sign (-1) and $\geq$ with a positive sign $(+1)$, such that a *signed feature* can be written as a tuple $(k, b_k) \in [p] \times \{-1, +1\}$. Also, although we assume $|Y| < 1$, the constant 1 does not matter here as we can scale Y by any constant and the conclusions in our main Theorem 9.4.1 below will still hold.

Remark 1: Because RF remains invariant under any strictly monotone transform of an individual feature, our results still hold when the uniform distribution assumption of X in the

LSS model is relaxed to the assumption that individual features X_j, $j \in [p]$, are independent with a distribution that has Lebesgue density.

Remark 2: The assumption in the LSS model that different interactions S_{j_1}, S_{j_2} with $j_1 \neq j_2$ are disjoint (see the (LSS-expectation) condition) is also probably not justified in many real data applications. The general problem with overlapping interactions in the LSS model is that such models can be non-identifiable, meaning that different forms of (9.1) can imply the same regression function $E(Y|X)$. For example, for the response $\mathbf{1}(X_1 < 0.5, X_2 < 0.5) + \mathbf{1}(X_1 > 0.5, X_2 > 0.5)$, by the definition of signed interactions in Definition 9.1.1, it has two basic signed interactions $\{(1,-1),(2,-1)\}$ and $\{(1,+1),(2,+1)\}$. However, we can also write it as $1 - \mathbf{1}(X_1 < 0.5, X_2 > 0.5) - \mathbf{1}(X_1 > 0.5, X_2 < 0.5)$, which has two different basic interactions $\{(1,-1),(2,+1)\}$ and $\{(1,+1),(2,-1)\}$. This means, a set of features which is an interaction in one of the representations is not an interaction in the other. Due to this identifiability problem, overlapping features can lead to both false positives and false negatives in term of interaction recovery with RF. One may try to define interaction more broadly to avoid this identifiability problem. For the previous example $\mathbf{1}(X_1 < 0.5, X_2 < 0.5) + \mathbf{1}(X_1 > 0.5, X_2 > 0.5)$, although the basic interactions are not unique, they always constitutes of both X_1 and X_2. Whether the coefficients $\{\beta_j\}_{j=0}^J$ are allowed to have different signs also affects the identifiability. The previous example is identifiable if we only allow positive coefficients. For a particular application, one should investigate identifiability further, but as this depends on the precise application, we leave this for future work. Our work provides the pathway to analyze this in detail.

The LSS model is trying to capture interactive thresholding behavior which has been observed for various biological processes [98, 26, 52, 42, 51, 49]. For example, in gene regulatory networks often a few different expression patterns are possible. Switching between those patterns can be associated with individual components that interact via a threshold effect [52, 42, 51]. Such a threshold behavior is also observed for other signal transduction mechanisms in cells, e.g, protein kinase [26] and cell differentiation [98]. Another example of a well studied threshold effect is gene expression regulation via small RNA (sRNA) [49]. Although for most biological processes the precise functional mechanisms between different features and a response variable of interest are much more complicated than what the LSS model can capture, theoretical investigations of a particular learning algorithm, such as RF, are only feasible within a well defined and relatively simple mathematical model. Given the empirically observed interactive threshold effects in many real biological systems, the LSS model clearly provides an enrichment to the current state of affairs, since current theoretical models do not capture the often observed interactive threshold behavior.

In the following we show that the RF algorithm can recover basic interactions $S_1, \ldots, S_J$ in the LSS model. Besides recovering $S_j \subset [p]$, RF can also recover the signs of each feature $k \in \cup_{j=1}^J S_j$ in the LSS model, which indicates whether the corresponding threshold behavior in (9.1) is given by a $\leq$- or a $\geq$-inequality. Without loss of generality, in the rest of the thesis

we assume that all inequalities are $\leq$ in (9.1), that is,

$$E(Y|X) = \beta_0 + \sum_{j=1}^{J} \beta_j \prod_{k \in S_j} \mathbf{1}(X_k \leq \gamma_k). \tag{9.2}$$

We stress, however, that all our results also hold for the general case (9.1). We call the unsigned set of features $S_1, \ldots, S_J \subset [p]$ the *unsigned basic interactions* of the LSS model and call the signed set of features $S_1^-, \ldots, S_J^- \subset [p] \times \{-1, +1\}$ with $S_j^- = \{(k, -1) \ : \ k \in S_j\}$ the *signed basic interactions* of the LSS model. As our theoretical results will show, RF does not just recover the unsigned interactions $S_j \subset [p]$, but also signed interactions $S_j^- \subset [p] \times \{-1, +1\}$. In other words, RF not only recover which features interact with each other in a LSS model, but also recover whether a particular feature in an interaction has to be larger or smaller than some threshold for this interaction to be active. Besides the signed and unsigned *basic interactions* we also define a *union signed interaction* as the union of individual basic signed interaction, as made more precise in the following definition.

Definition 9.1.1. *In the LSS model with basic signed interactions* $S_1^-, \ldots, S_J^- \subset [p] \times \{-1, +1\}$ *a (non-empty) set of signed features* $S^{\pm} \subset [p] \times \{-1, +1\}$ *is called a union signed interaction, if*

$$S^{\pm} = \bigcup_{j \in \mathcal{I}} S_j^- \bigcup_{j \in \mathcal{I}_s, k \in S_j} \{(k, b_k) \ : \ b_k \in \{-1, +1\}\} \tag{9.3}$$

for some (possibly empty) set of indices $\mathcal{I} \subset \{j \in [J] \ : \ |S_j| > 1\}$, $\mathcal{I}_s \subset \{j \in [J] \ : \ |S_j| = 1\}$.

For interactions with only one feature k, due to the sign ambiguity in the LLS model, i.e., $\mathbf{1}(X_k \leq a) = 1 - \mathbf{1}(X_k > a)$, both, $(k, -1)$ and $(k, +1)$, will be counted as an interaction.

The theoretical results, that we present in Section 9.4 are asymptotic, in the sense that they assume the sample size n to go to infinity. The number of signal features in the LSS model is assumed to be bounded by s (independent of n and p). However, the overall number of features p or the number of noisy features $p - s$ can grow to infinity as n increases. Mathematically, our theoretical results assume A1.

A1 (Sparsity)**.** $s = O(1)$ *and* $\frac{\log(p)}{n} \to 0$.

This means that, in contrast to many theoretical works [22, 83, 94], our results hold in a high-dimensional setting, as long as the overall number of signal features s is bounded. See also [11] for results that only depend on s and not p and thus, cover the high-dimensional setting, too.

9.2 Technical assumptions and notations

RF is an ensemble of classification and regression trees, where each tree T defines a mapping from the feature space to the response. Trees are constructed independently of one another

on a bootstrapped or subsampled data set $\mathcal{D}^{(T)}$ of the original data $\mathcal{D}$. Any node t in a tree T represents a hyper-rectangle R_t in the feature space. A split of the node t is a pair (k_t, γ_t) which divides the hyper-rectangle R_t into two hyper-rectangles $R_{t,l}(k_t, \gamma_t) = R_t \cap \mathbf{1}(X_{k_t} \leq \gamma_t)$ and $R_{t,r}(k_t, \gamma_t) = R_t \cap \mathbf{1}(X_{k_t} > \gamma_t)$, corresponding to the left child t_l and right child t_r of node t, respectively. For a node t in a tree T, $N_n(t) = |\{i \in \mathcal{D}^{(T)} : \mathbf{x}_i \in R_t\}|$ denotes the number of samples falling into R_t.

Each tree T is grown using a recursive procedure which proceeds in two steps for each node t. First, a subset $M_{\text{try}} \subset [p]$ of features is chosen uniformly at random. The size of M_{try} is m_{try}. Then the optimal split $k_t \in M_{\text{try}}, \gamma_t \in \mathbb{R}$ is determined by maximizing impurity decrease defined in (9.4):

$$\Delta_I^n(t) := I_n(t) - \frac{N_n(t_l)}{N_n(t)} I_n(t_l) - \frac{N_n(t_r)}{N_n(t)} I_n(t_r) \tag{9.4}$$

where $t_l(t_r)$ is the left(right) child of t and $I_n(t)$ is an impurity measure. Here, $I_n(t)$ is defined as the variance of the response y_i's for all the samples in the region R_t. The procedure terminates at a node t if two children contain too few samples, e.g., $\min\{N_n(t_l), N_n(t_r)\} \leq 1$, or if all responses are identical. For any tree T and any leaf node $t_{\text{leaf}} \in T$, denote $\mathfrak{p}(t_{\text{leaf}})$ to be a path to that leaf node and $D(\mathfrak{p}(t_{\text{leaf}}))$ to be its depth. For any hyper-rectangle R_t, $\mu(R_t)$ denotes its volume. We have the following assumptions on RF:

A2 (increasing depth)**.** *The minimum depth of any path in any tree goes to infinity, i.e.,* $\min_T \min_{t_{leaf} \in T} D(t_{leaf}) \xrightarrow{p} \infty$ *as* $n \to \infty$*.*

A3 (balanced split)**.** *Each split* (k_t, γ_t) *is balanced: for any node* t*,*

$$\min\left(\frac{\mu(R_{t,l}(k_t, \gamma_t))}{\mu(R_{t,r}(k_t, \gamma_t))}, \frac{\mu(R_{t,r}(k_t, \gamma_t))}{\mu(R_{t,l}(k_t, \gamma_t))}\right) > \frac{C_\gamma}{1 - C_\gamma}.$$

A4 (m_{try})**.** $C_m p + (1 - C_m)s \leq m_{try} \leq (1 - C_m)(p - s)$ *where* $C_m \in (0, 1)$ *is a constant.*

A5 (No bootstrap)**.** *All the trees in RF are grown on the whole data set without bootstrapping, i.e.* $\mathcal{D}^{(T)} = \mathcal{D}$ *for any* T*.*

A2 ensures that the length of any decision path in any tree tends to infinity. This assumption is reasonable as tree depths in RF is usually of order $O(\log n)$ which tends to infinity as $n \to \infty$. A3 ensures that each node split is balanced. Similar conditions are used commonly in other papers [94]. A4 shows the important role of the parameter m_{try}. Roughly speaking, m_{try} cannot be too small or too big. When m_{try} is too small, there will be too many splits on irrelevant features which makes the tree noisy. When m_{try} is too big, there will be too little variability in the tree structure. This motivation will be made rigorous in the proof. A5 is a technical assumption to simplify our analysis. Since we study the asymptotic case, bootstrap has little impact on the tree structure, which means it will not affect our result.

9.3 Depth weighted prevalence (DWP)

Given a tree T in RF, we can randomly select a path $\mathcal{P}$ of T as follows: we start at the root node of T and then, at every node, randomly go left or right until we reach a leaf node. This is equivalent to selecting a path in T of depth D with probability 2^{-D}. As such, any path $\mathcal{P}$ in a decision tree can be associated with a sequence of signed features $(k_1, b_{k_1}), \ldots, (k_D, b_{k_D}) \in [p] \times \{-1, +1\}$, where D is the depth of the path and for any node $t = [D]$ on the path the sign b_{k_t} indicates whether the path at node t followed the $\leq$ direction ($b_{k_t} = -1$) or the $>$ direction ($b_{k_t} = +1$) for the split on feature $k_t \in [p]$. For the randomly selected path $\mathcal{P}$ and any fixed constant $\epsilon > 0$, we now define $\hat{\mathcal{F}}_\epsilon(\mathcal{P}, \mathcal{D})$ to be the set of signed features on $\mathcal{P}$ where the corresponding node in the RF had an impurity decrease of at least ϵ, that is,

$$\begin{aligned}\hat{\mathcal{F}}_\epsilon(\mathcal{P}, \mathcal{D}) :=& \{(k_t, b_{k_t}) \mid \ t \text{ is a node of } \mathcal{P} \text{ with } \Delta_I^n(t) > \epsilon \text{ and} \\ & k_t \text{ appears first time on } \mathcal{P}\}. \end{aligned} \tag{9.5}$$

We use $\hat{\mathcal{F}}_\epsilon$ as a shorthand for $\hat{\mathcal{F}}_\epsilon(\mathcal{P}, \mathcal{D})$ when the path $\mathcal{P}$ and the data $\mathcal{D}$ of interest are clear. Note that if a feature appears more than once on the path $\mathcal{P}$, its sign in $\hat{\mathcal{F}}_\epsilon(\mathcal{P})$ is the sign when the feature appears the first time with the impurity decrease above the threshold. Our main theorems will be stated in terms of the prevalence of signed feature set $S^\pm \subset [p] \times \{-1, +1\}$ on the random path $\mathcal{P}$ within $\hat{\mathcal{F}}_\epsilon(\mathcal{P})$, where $\mathcal{P}$ is a random path. To formally define the prevalence of $S^\pm$, we first need to identify the sources of randomness underlying the random path $\mathcal{P}$. There are three layers of randomness involved:

($\mathcal{D}$: Data randomness) the randomness involved in the data generation;

(T: Tree randomness) the randomness involved in growing an individual tree with parameter m_{try}, given data $\mathcal{D}$;

($\mathcal{P}$: Path randomness) the randomness involved in selecting a random path $\mathcal{P}$ of depth d with probability 2^{-d}, given the tree T.

In our following definition of the prevalence of signed feature sets, the probability is conditioned on the data $\mathcal{D}$, and taken only over the randomness of the tree T and the randomness of selecting one of its path as in $\mathcal{P}$.

Definition 9.3.1. *(Depth-Weighted Prevalence (DWP)) For any signed feature set $S^\pm \subset [p] \times \{-1, +1\}$, conditioned on data $\mathcal{D}$, we define the* Depth-Weighted Prevalence (DWP) *of $S^\pm$ as the probability that $S^\pm$ appears on the random path $\mathcal{P}$ within the set $\hat{\mathcal{F}}_\epsilon$, that is,*

$$\text{DWP}(S^\pm) = P_{(\mathcal{P},T)}(S^\pm \subset \hat{\mathcal{F}}_\epsilon(\mathcal{P}) \mid \mathcal{D}). \tag{9.6}$$

While we only have a fixed sample size which means the data randomness is inevitable, the tree randomness and path randomness are generated by the algorithm and thus can be eliminated by sampling as many trees and paths as we like. Because the depth-weighted prevalence in (9.6) is only conditioned on the data, for any given $\epsilon > 0$ and set of signed features $S^\pm$, it can be computed with arbitrary precision.

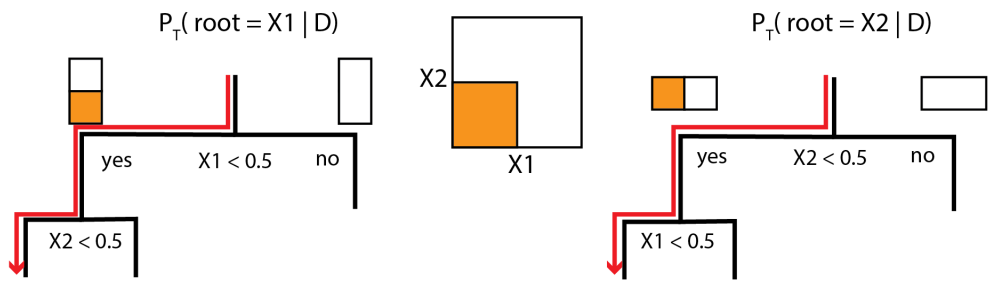


Figure 9.1: Illustration of RF trained using data from (9.7)

9.4 Main results

Before we state our main results in full detail, we want to illustrate it with a simple example.

Illustrative example: Assume that $p = 2$ and there are just two features X_1 and X_2. Assume there is a single interaction $J = 1$ and the regression function is (9.7).

$$E(Y|X_1, X_2) = \mathbf{1}(X_1 \leq 0.5) \cdot \mathbf{1}(X_2 \leq 0.5). \tag{9.7}$$

The response surface of (9.7) is shown in Figure 9.1 in the top middle plot. We consider the population case, where we have full access to the joint distribution $P(X, Y)$, that is, we have access to an unlimited amount of data ($n = \infty$). When we apply the RF algorithm as in Section 9.2, then for each individual tree in the forest the root node either splits on feature X_1 or on feature X_2. Since X_1 and X_2 are completely symmetric in the distribution $P(X, Y)$, thus, if the RF algorithm grows infinitely many trees, half of them will split on X_1 at the root node and half of them split on X_2 at the root node. Furthermore, as the split threshold for every node in the tree maximizes impurity decrease, the split will be at 0.5 for any feature. This is illustrated in Figure 9.1, where the left bottom figure shows a tree which splits on feature X_1 at the root node and the right bottom figure shows a tree which splits on feature X_2 at the root node. As each tree in RF grows to purity, when the root node splits at feature X_1, then for the path of the tree which follows the $(1, +1)$ direction, the tree will stop growing, as the respective response surface is already constant. However, for the path of the tree which follows the $(1, -1)$ direction, the tree will further split on the remaining feature X_2. Thus, we conclude that the forest consists of exactly the two different trees shown in Figure 9.1, each appearing equally often. For each node t in these trees, the impurity decrease $\Delta_I^n(t) \geq 1/16$. Thus, for any $\epsilon < 1/16$, we can write the DWP of the basic

signed interaction $S^- = \{(1,-1),(2,-1)\}$:

$$\begin{aligned}&\mathrm{DWP}(S^-) = \\ &\underbrace{P_T(T\text{'s root splits on feature 1})}_{=0.5}\cdot 2^{-2} + \\ &\underbrace{P_T(T\text{'s root splits on feature 2})}_{=0.5}\cdot 2^{-2} = 2^{-2} = 2^{-|S^-|}.\end{aligned}$$

In Figure 9.1 the paths which contain the basic signed interaction $S^- = \{(1,-1),(2,-1)\}$ are marked red. For all the other sets of signed features $S^\pm \subset [p]\times\{-1,+1\}$, it can be shown that

$$\mathrm{DWP}(S^\pm) < 2^{-|S^\pm|}.$$

For example,

$$\mathrm{DWP}(\{(1,-1),(2,+1)\}) = 2^{-3} < 2^{-2}$$

and

$$\mathrm{DWP}(\{(1,-1)\}) = 2^{-2} + 2^{-3} < 2^{-1}.$$

As we will formally state in the theorem below, the same reasoning holds true asymptotically for any RF trained on the data from the LSS model, namely, the DWP of a set of signed features $S^\pm \subset [p]\times\{-1,+1\}$ is always upper bounded by $2^{-|S^\pm|}$ and this upper bound is attained if and only if $S^\pm$ is a union signed interaction.

Theorem 9.4.1. *For any impurity threshold $\epsilon > 0$, let*

$$\tilde{\epsilon} := \left(4\epsilon/(C_\beta^2 C_\gamma^{2s-1})\right)^{C_m^{2s}/\log(1/C_\gamma)}.$$

For any set of signed features $S^\pm \subset [p]\times\{-1,+1\}$, conditioned on some input data $\mathcal{D}$ for the RF algorithm from Section 9.2, it holds true that

1. *(General upper bound)*
$$\mathrm{DWP}(S^\pm) \le 2^{-|S^\pm|},$$
2. *(Interaction lower bound) When $S^\pm$ is a union interaction as in Definition 9.1.1, then,*
$$\mathrm{DWP}(S^\pm) \ge 2^{-|S^\pm|} - \tilde{\epsilon} - r_n(\mathcal{D},\epsilon),$$
3. *(Non-interaction upper bound) when $S^\pm$ is not a union interaction as in Definition 9.1.1, then,*
$$\mathrm{DWP}(S^\pm) \le 0.5^{|S^\pm|}\left(1 - 0.5\, C_m^s\right) + r_n(\mathcal{D},\epsilon).$$
Under the LSS model, for every fixed $\epsilon > 0$ it holds that
$$r_n(\mathcal{D},\epsilon) \xrightarrow{p} 0 \quad \text{as } n\to\infty,$$
where $\xrightarrow{p}$ denotes convergence in probability.

Proof Sketch: The detailed proof of Theorem 9.4.1 has two major parts: first, showing the assertion for the idealized population case and second, extending the population case to the finite sample case. The major difficulty of the first step is to define an adequate *population version* of the set $\hat{\mathcal{F}}_\epsilon$. To this end, in the proof in the SI we define a set $\mathcal{F}$, which we denote *desirable features*, which correspond to all features of a path $\mathcal{P}$ which would result in a decrease in impurity if the RF would get to see the full distribution $P(X, Y)$ (not just a finite sample $\mathcal{D}$) and thus, would split at the exact thresholds γ. The set of *desirable features*, $\mathcal{F}$, is an oracle, in the sense that its construction depends on the true underlying LSS model. This is in contrast to the set $\hat{\mathcal{F}}_\epsilon$, which can be computed for any given RF. Given this definition of *desirable features*, a sketch of the proof of the major assertions of Theorem 9.4.1 is as follow: When a set of signed features $S^\pm$ appears on $\mathcal{F}$, this implies that every time a features $(k, b_k) \in S^\pm$ appeared on the way from the root node to the leaf, the correct splitting direction was selected for $\mathcal{P}$, which gives rise to the general upper bound of $\mathrm{DWP}(S^\pm) \leq 2^{-|S^\pm|}$. If $S^\pm$ is a union interaction, then (assuming all paths of the tree are pure) a correct splitting direction for each of its features already implies that $S^\pm$ appears on $\mathcal{P}$ and thus, $\mathrm{DWP}(S^\pm) = 2^{-|S^\pm|}$. If $S^\pm$ is not a union interaction, then there will always be the possibility that, although every split for an encountered feature which is an element of $S^\pm$ was done in the correct direction, some of the features in $S^\pm$ were just never encountered and therefore, a correct splitting direction does not imply that $S^\pm$ appears on $\mathcal{P}$, hence $\mathrm{DWP}(S^\pm) < 2^{-|S^\pm|}$. In the second step of the proof, we show that the observed set $\hat{\mathcal{F}}_\epsilon$ and the oracle set $\mathcal{F}$ are the same with high probability. In order to prove that, we need to show the trees grown using finite samples are close to the trees grown using the population in terms of the splitting features and thresholds as long as the feature is desirable. The the tricky part is that a tree grown using finite samples can deviate from a tree grown using the population when a node splits on noisy features. Thus, we need to carefully analyze these two cases separately. This part of the proof mainly rely on uniform convergence results.

Remark 3: This theorem relies on the assumption that Y is bounded. If we assume a slightly stronger assumption on p and n than A1: $(\log n)^{1+\delta} \log p/n \to 0$ for some $\delta > 0$, then the conclusions still hold when the noise $Z := Y - E(Y|X)$ is independent of X and 1-subgaussian, that is,

$$E(\exp(tZ)) \leq \exp(t^2/2) \text{ for all } t \in \mathbb{R}.$$

See Proposition B.2.2 in the appendix for more detail.

Remark 4: Our theory shows that recovery of interactions becomes exponentially more difficult as size of interaction increases – therefore one should only aim to recover small to moderately sized interactions. An interaction of size s correspond to an region of size $O(2^{-s})$, which means the sample size must be much larger than 2^s to have enough sample in that region. Also, the DWP of a basic interaction of size s is 2^{-s}. To have a consistent estimate, the number of independent paths should be much larger than 2^s. Thus, when one wants to recover an interaction of size s, the number of samples and the number of trees must be much larger than 2^s. That shows the intrinsic difficulty of estimating high order interactions.

Recall that the parameter ϵ in Theorem 9.4.1 can, in principle, be chosen arbitrarily small,

and thus, $\tilde{\epsilon}$ in Theorem 9.4.1 can be made arbitrarily small. Hence, up to an arbitrarily high precision, Theorem 9.4.1 implies that asymptotically union interactions are exactly those set of signed features $S^{\pm}$ whose DWP, i.e., $\text{DWP}(S^{\pm})$, attains the upper bound $2^{-|S^{\pm}|}$. Recall that the DWP is computable from data. Hence, if one had access to (upper bounds for) the constants C_β, C_γ, and s (recall that C_m is known, as m_{try} is known), one could select ϵ small enough such that

$$\tilde{\epsilon} < (C_m/2)^s \, .$$

Then, if follows from Theorem 9.4.1, that the algorithm which classifies $S^{\pm}$ as a union interaction if and only if $\text{DWP}(S^{\pm}) \geq 2^{-|S^{\pm}|} - (C_m/2)^s$ is consistent (i.e., classifies correctly as $n \to \infty$) under the LSS model. This shows that the RF algorithm via its DWP consistently recovers union interactions, whenever ϵ is chosen sufficiently small.

Note that any algorithm which can consistently recover *union signed interactions*, can also consistently recover the *signed basic interactions* of the LSS model, as those are just the smallest unites within the set of all union interactions. The only exception is for a signed basic interaction of size one, for which the sign is not identifiable from the LSS model.

Remark 5: One important assumption in our theorem is the sparsity of signal features. If there are many "weak" signal features, it is very hard for RF to work well. For RF, at each node of a tree, only one feature is used. That means the total number of features used along each path is limited by the depth of the tree, which is usually of order $O(\log n)$. For our assertions of Theorem 9.4.1 the hard threshold ϵ in the set $\hat{\mathcal{F}}_\epsilon$ has the purpose to select the signal features. Clearly, the choice of an appropriate value of ϵ is hard in practice, as an optimal choice depends on the LSS model (recall the previous paragraph). The iterative random forest fitting procedure in iRF [10] (which uses joint prevalence on decision paths in RF to recover interactions, similar as suggested by Theorem 9.4.1) filters noisy features not with a hard, but with a soft thresholding procedure: it grows several RF iteratively and samples features at each node according to their feature importance from the previous iteration. In that way, one does not need to chose a single hard threshold. We follow this strategy for our simulation results.

One of the remarkable aspects of the results in Theorem 9.4.1 is that the DWP of a signed interaction is asymptotically independent of any model coefficients. That is, it only depends on its size $|S^{\pm}|$ and nothing else. In a sense, this shows that the tree structure of RF contains the *qualitative* information of *which features interact with each other*, independently of the *quantitative* information about *what are the precise parameters* in the LSS model.

Another interesting aspect about the results from Theorem 9.4.1 is that it sheds some light on the influence of m_{try} on the the interaction recovery performance of RF. For the third assertion in Theorem 9.4.1 we actually show that $\text{DWP}(S^{\pm}) \leq r_n(\mathcal{D}, \epsilon) +$

$$0.5^{|S^{\pm}|} \left(1 - 0.5 \min_{k \in \cup_j S_j} P(\text{root node splits on feature } k) \right) .$$

When m_{try} is too large,

$$\min_{k \in \cup_j S_j} P(\text{root node splits on feature } k)$$

can get very small, as particularly strong features (large initial impurity decrease) can mask weaker features. As an extreme example, consider the situation where $m_{try} = p$ and thus, the root node gets to see all the features. In that case, the single feature which has the highest impurity decrease, say X_1, will *always* appear at the root node and hence, for $S^{\pm} = \{(1, -1)\}$ or $S^{\pm} = \{(1, +1)\}$ one will get $\mathrm{DWP}(S^{\pm}) = 2^{-|S^{\pm}|} = 0.5$, independent of whether $S^{\pm}$ is an interaction or not. This shows that when m_{try} is too large, false interactions' DWP can attain the universal upper bound $2^{-|S^{\pm}|}$, which leads to false positives in terms of interaction recovery. On the other hand, when m_{try} is too small, for a signal feature $k \in \cup_j S_j$ it can take a long time until it gets selected into the candidate feature set at a node. In particular, for finite sample, it can happen that the tree reaches purity due to lack of samples without having split on any of the signal features. Hence, the reasoning of Theorem (9.4.1), namely that *correct split direction + pure path* implies that a union interaction appears on the path does not hold anymore. This can lead to union interactions having significantly smaller DWP than the universal upper bound $2^{-|S^{\pm}|}$, i.e., false negatives in terms of interaction recovery.

Chapter 10

LSSrank and simulation results

10.1 LSSrank: a theoretically inspired ranking criterion for boolean interactions

To conduct follow-up experiments, biologists need to know which interactions are most promising candidates to follow up on in experiments, e.g., they want to know *what are the three most promising candidates and in which order*. In the following, we build on the results from Theorem 9.4.1 to propose a ranking criterion for candidate interactions: Under the LSS model, for large sample size ($n \to \infty$) and ϵ sufficiently small, our theoretical results from Section 9.4 show that for any set of signed features $S^{\pm} \subset [p] \times \{-1, +1\}$ of size $|S^{\pm}|$ with RF depth weighted prevalence $\mathrm{DWP}(S^{\pm})$, it holds true that

$$S^{\pm} \text{ is a union interaction} \iff \frac{\log_2(\mathrm{DWP}(S^{\pm}))}{|S^{\pm}|} \approx -1,$$
$$S^{\pm} \text{ is not a union interaction} \iff \frac{\log_2(\mathrm{DWP}(S^{\pm}))}{|S^{\pm}|} < -1.$$

Thus, for given sets of signed candidate interactions $S_1^{\pm}, \ldots, S_M^{\pm}$, our theoretically inspired **LSSrank** criterion ranks the candidates in a decreasing order based on

$$\rho(S^{\pm}) := \frac{\log_2(\mathrm{DWP}(S^{\pm}))}{|S^{\pm}|},$$

that is, the interaction $S^{\pm}$ with largest $\rho(S^{\pm})$ comes first in the ranking.

Remark 6: Recall that our main Theorem 9.4.1 only holds for the asymptotic case ($n \to \infty$). As we noted in Remark 4, the number of samples needed to observe an interaction of order s must grow exponentially with s. Moreover, as the DWP is upper bounded by $2^{-|S|}$ is also follows that the number of paths (and trees) that need to be generated in order calculate DWP with a sufficient precision also has to grow exponentially with s. Therefore, we stress that applying LSSrank is only reasonable for interactions of moderate size.

Illustrative example, Here we illustrate the idea of LSSrank via a simple example: $Y = \mathbf{1}(X_1 \leq 0.5, X_2 \leq 0.5) + \mathbf{1}(X_3 \leq 0.5, X_4 \leq 0.5)$. We simulate 5000 samples. Then we use RF with 200 trees. Here is the ρ-value for a few signed interactions: $\rho(\{(1,-1),(2,-1)\}) = -1$, $\rho(\{(1,+1),(2,-1)\}) = -1.3$, $\rho(\{(3,-1)\}) = -1.27$, $\rho(\{(1,-1),(2,-1),(3,-1)\}) = -1.1$. As can be seen, the basic interaction has the highest value.

There are many ways to obtain a **list of candidate interactions** from data. The naive way is via brute force: computing all combinations of features. However, this is not computationally feasible for a moderate number of features. Since we only care for the interactions with high prevalence, we can use the FP-growth algorithm or Random Interaction Trees (RIT) to obtain the interactions with prevalence higher than some threshold. While RIT computes the candidate interactions faster, it is a probabilistic algorithm so some candidate interactions with high prevalence might be missing. Thus, here we use FP-growth, which runs in reasonable time and gives more robust results than RIT.

10.2 Simulated data from LSS models

We first simulate the data from an LSS model with different number of basic interactions and interaction orders. Set $p = 50$ and $n = 1,000$. Each feature X_j is generated from an uniform distribution $U([0,1])$, independent from one another. The number of basic interactions is denoted as K; the order of each interaction by L; we consider the same threshold τ for all features; and additive Gaussian noise with variance σ^2, then the response is:

$$Y = \sum_{k=1}^{K} \prod_{\ell=(k-1)\cdot L+1}^{k\cdot L} \mathbf{1}(X_\ell < \tau) + \mathcal{N}(0, \sigma^2).$$

We consider a variety of values for K, L and σ^2, namely, $K = 2, \ldots, 5$, $L = 2, \ldots, 4$, and σ^2s such that the signal-to-noise ratios (SNR) is given by $1, 20, 50$ or 100. For a given K and L, the threshold τ is chosen such that about 50 percent of samples fall into the union of hyper-rectangles, that is, $\cup_{k=1}^{K} \cap_{\ell=(k-1)\cdot L+1}^{k\cdot L} \{X_\ell < \tau\}$. This criterion roughly ensures that the information about the basic interactions in X is comparable across different number of interactions and interaction orders. The results are averaged across 40 independent Monte Carlo runs. In order to apply LSSrank, we first need to grow a RF. Recall that our results also assume an MDI threshold for the individual nodes of the tree. Here, we follow a soft thresholdind strategy via iRF to implement this. We use 10 iterations and each time grow 300 trees. Given a basic interaction $S^* \subset [p]$ and an estimated interaction $\hat{S} \subset [p]$, we evaluate their proximity based on their Jaccard distance:

$$\text{score}(S^*, \hat{S}) = \frac{|S^* \cap \hat{S}|}{|S^* \cup \hat{S}|}.$$

Given the K true basic interactions $S_1^*, \ldots, S_K^*$ from the respective LSS model and the top M interactions from the estimated LSSrank ranking $\hat{S}_1, \ldots, \hat{S}_M$, we define their proximity

score to be

$$\text{score}(\{S_j^*\}_{j=1}^K, \{\hat{S}_i\}_{i=1}^M) = \frac{1}{K}\sum_{j=1}^{K} \max_{i=1}^{M} \text{score}(S_j^*, \hat{S}_i). \tag{10.1}$$

In other words, for each basic interaction S_j^*, we find the estimated interaction that has the highest proximity, and then compute the average scores across all S_j^*. Note that this score will increase monotonically as M increases. We also note that, although our LSSrank score is based on signed features, our evaluation criterion is based on the respective unsigned features. We choose not to use the signed features because in certain simulations, one feature have different signs in different basic interactions, which makes the meaning of the sign of a feature ambiguous. The simulation results are shown in Fig. 10.1. For example, the plot in the 1st row, 1st column shows that LSSrank's top interaction ($M = 1$) corresponds to the basic interaction in all Monte Carlo runs and among all the considered SNRs (proximity score = 1). Similar, the plot in the 2nd row, 2nd column shows that the top two interactions ($M = 2$) correspond to the two basic interactions in all the Monte Carlo runs when $SNR = 100$. However, for weaker signals with $SNR = 1$ the proximity score is only around 0.75 and thus, in some situations the top two interactions do not coincide with the two basic interactions. In general, the performance gradually degrades when the number of basic interactions and the order of interactions increases. Note that this is consistent with our theoretical results, where constants in the $o(1)$ terms depend on K and L, see Theorem 9.4.1. We also note that, when SNR is higher than 10, the scores do not change much, which indicates that LSSrank is robust to highly noisy responses.

10.3 Robustness to LSS model violations:

We investigate the stability of results of our method when the model assumptions are not met. We consider two different settings, both with $n = 1000$ and $p = 50$. First, we consider $SNR = 50$, with 3 order-3 interactions, analog as in Figure 10.1, 3rd row, 2nd column, green line. Second, we consider $SNR = 100$, with 2 order-3 interactions, analog as in Figure 10.1, 2nd row, 2nd column, red line. We consider a variety of perturbations:

- **Overlapping interactions:** different basic interactions have overlapping features. When overlap $= k$, the basic interactions are $((1, -1), (2, -1), (3, -1))$, $((4 - k, -1)$, $(5 - k, -1)$, $(6 - k, -1))$, and $((7 - 2k, -1)$, $(8 - 2k, -1)$, $(9 - 2k, -1))$ when $K = L = 3$ and $((1, -1)$, $(2, -1)$, $(3, -1))$, $((4 - k, -1)$, $(5 - k, -1)$, $(6 - k, -1))$ when $K = 2$ and $L = 3$.
- **Correlated features:** different features are correlated instead of independent. When corr $= \alpha$, the correlation between feature j_1 and j_2 is $\alpha^{|j_1 - j_2|}$.
- **Heavy-tail noise:** the noise follow Laplace or Cauchy distributions which has heavier tails than (sub-)Gaussian distributions.

Results are shown in Figure 10.2. For heavy tail noise (see bottom plot) we observe almost no drop in performance compared to Figure 10.1. For the correlated case, one can see a significant drop in performance, which matches well to our findings for the real data example (see next section). Two (false) interaction that often appears on top of LSSrank's list for the correlated case, are the order-one interaction $\{(2,+1)\}$and $\{(2,-1)\}$. This may be explained by the fact that the X_2 feature is strongly correlated with both, the X_1 and the X_3 feature. Thus, the signed interaction $\{(1,-1),(2,-1),(3,-1)\}$ partially merges into the order-one signed interaction $\{(2,-1)\}$). For the overlapping case, the situation is more complex. When $K=2$ and $L=3$, one observes a slight drop in performance. However, when $K=L=3$ the performance even improves in the overlapping case. This can be explained by to competing effects: One the one hand, for the overlapping case the overall number of signal features decreases, which generally leads to an increase in performance. On the other hand, the model violations in the overlapping case can lead to a decrease in performance.[1]

10.4 Real-data inspired simulations

We reconsider the enhancer data set that was already analyzed in [45, 10]. Following the general setup as in [45], from the corresponding feature matrix we simulate binary responses as follows:

- Single-component AND rule (AND)

$$\begin{aligned} &P(Y=1|X) = \\ &0.8 \cdot \mathbf{1}_{X_1 \geq q_{1,1-\alpha}, X_2 \geq q_{2,1-\alpha}, X_3 \geq q_{3,1-\alpha}, X_4 \geq q_{4,1-\alpha}}. \end{aligned} \tag{10.2}$$

- Multi-component AND rule (OR)

$$\begin{aligned} &P(Y=1|X) = \\ &0.8 \cdot \big[\mathbf{1}_{X_1 \leq q_{1,\alpha}, X_2 \leq q_{2,\alpha}, X_3 \geq q_{3,1-\alpha}, X_4 \geq q_{4,1-\alpha}} \\ &\text{ or } \mathbf{1}_{X_1 \geq q_{1,1-\alpha}, X_2 \geq q_{1,1-\alpha}, X_3 \leq q_{1,\alpha}, X_4 \leq q_{1,\alpha}}\big]. \end{aligned} \tag{10.3}$$

- Additive AND rule (ADD)

$$\begin{aligned} &P(Y=1|X) = \\ &0.4 \cdot \big[\mathbf{1}_{X_1 \geq q_{1,1-\alpha}, X_2 \geq q_{2,1-\alpha}, X_3 \geq q_{3,1-\alpha}} \\ &+ \mathbf{1}_{X_4 \geq q_{4,1-\alpha}, X_5 \geq q_{5,1-\alpha}, X_6 \geq q_{6,1-\alpha}}\big]. \end{aligned} \tag{10.4}$$

[1] When $K=2$ and $L=3$, one false interaction which often appears on top of LSSrank's list is the order-1 interaction $\{(2,+1)\}$ (and $\{(2,-1)\}$, respectively). Partly, this can explained by the soft dimension reduction of iRF, where in the last iteration effectively only a few features remain. When m_{try} is larger than the total number of effective features, then the single overlapping feature X_2, which has strictly higher MDI than non-overlapping features, will consistently be split on at the root node and thus, appears on top of LSSrank, as it will have DWP $= 0.5$.

where $\alpha = 0.1^{1/4}$ and $q_{i,\alpha}$ refers to the left α-quantile of i-th feature. The choice of α is to mimic the class imbalance as in the original enhancer response. Thereby, in each Monte Carlo run the respective signal features $X_1, \ldots, X_6$ were chosen uniformly at random and responses Y_i for different observations $i \in [n]$ are independent conditioned on X. The results are shown in Figure 10.3 (blue line):

- (AND rule) First, consider the AND rule in (10.2), see left plot in Figure 10.3. One can check that here the SNR is roughly given by 1, so we can compare this with the 1st row, 1st column of Figure 10.1. We observe quite some significant drop in performance, although results are still much better than random guess.

- (ADD rule) Similar, for the ADD rule in (10.4), see right plot in Figure 10.3, we can compare with the 2nd row, 2nd column of Figure 10.1, where also a significant drop in performance is observed.

- (OR rule) The highest proximity score is observed for the OR rule in (10.3), see middle plot in Figure 10.3. This setting cannot directly be compared with Figure 10.1. One reason for the particularly high proximity score of the OR rule is it has effectively two different signed interactions which coincide as un-signed interactions. As the proximity score in (10.1) does not take the sign of the features into account, this results in a particularly high proximity score.

In order to better understand this drop in performance for the real data, we investigate two different sources of LSS model violations:

- (Correlation of enhancer features) The different features in the enhancer data set or quite strongly correlated, see Figure 10.4. In order to investigate the effect of this, we re-run the simulations, but with each of the features randomly permuted across samples. Results are shown in Figure 10.3 (orange line). As can be seen, the results slightly improve, although not much. Thus, we conclude that correlation is some, but not the major reason for the drop in performance.

- (Marginal feature distribution) We notes that a lot of features in the enhance data set have marginal distribution with a heavy point mass on zero (see, for example, the histogram of the Kr feature in Fig. 10.5). In order to investigate this effect, we further re-scaled the marginal distribution of each feature, such that it follows a Gaussian distribution, without any point-masses. Results are shown in Figure 10.3 (green line). As can be seen, removing the point mass on zero results in an almost perfect proximity score for LSSrank for all three models.

We conclude that performance of the LSSrank ranking criterion can be impeded in real data due to violations of the LSS model assumptions. For this particular enhancer data set, we found that the major source for decrease in performance seems to come from the marginal feature distributions which often has heavy point mass at zero.

10.5 Enhancer data

We used LSSrank to select the interactions for the real enhancer data and the results are similar to that of the original iRF paper. See Table 10.1 for a comparison between results from LSSrank and the results from original iRF [10]. The top-2 interactions of LLSrank and iRF are the same. Also, the top 20 interactions selected via LSSrank contains all the order-3 interactions and most order-2 interactions discovered by iRF.

Table 10.1: The top 20 interactions found by LSSrank and iRF. order-1 and order-2 interactions that are contained in any higher-order interactions are removed following similar treatments in Basu et al. [10]. An interaction is marked as blue if it appeared in both columns. An interaction is marked as orange if it strictly contains or is contained in an interaction in the other column. LSSrank discovered all the order-3 interactions that are discovered by iRF.

Top 20 interactions via LSSrank	Top 20 interactions via iRF
Gt_Kr_Twi	Zld_Gt_Twi
Zld_Gt_Twi	Gt_Kr_Twi
Gt_Med_Twi	Gt_Med
Zld_Gt_Kr	Gt_Hb
Zld_Gt_Kr_Twi	H3K36me3_Gt_Twi
H3k18ac_Gt_Twi	Bcd_Gt
Zld_Kr_Twi	Bcd_Twi
Gt_Kr_Med2_Twi	Med_Twi
H3k18ac_Gt_Kr_Twi	H3_Gt
Gt_Hb_Twi	H3K27me3_Gt
Gt_Hb_Kr	H3K27me3_Twi
H3k4me3_Gt_Twi	Hb_Kr
Gt_Kr_Med	H3K36me3_Zld
Bcd_Gt_Twi	H3K4me3_Gt_Twi
H3k18ac_Gt_Kr	H3K4me3_Kr
H3k9ac_Gt_Twi	Zld_Gt_Kr
H3k36me3_Gt_Twi	Hb_Twi
H3k27ac_Gt_Twi	H3K18ac_Kr
Kr_Med_Twi	Kr_Med
H3k4me1_Gt_Twi	H3K9ac_Kr

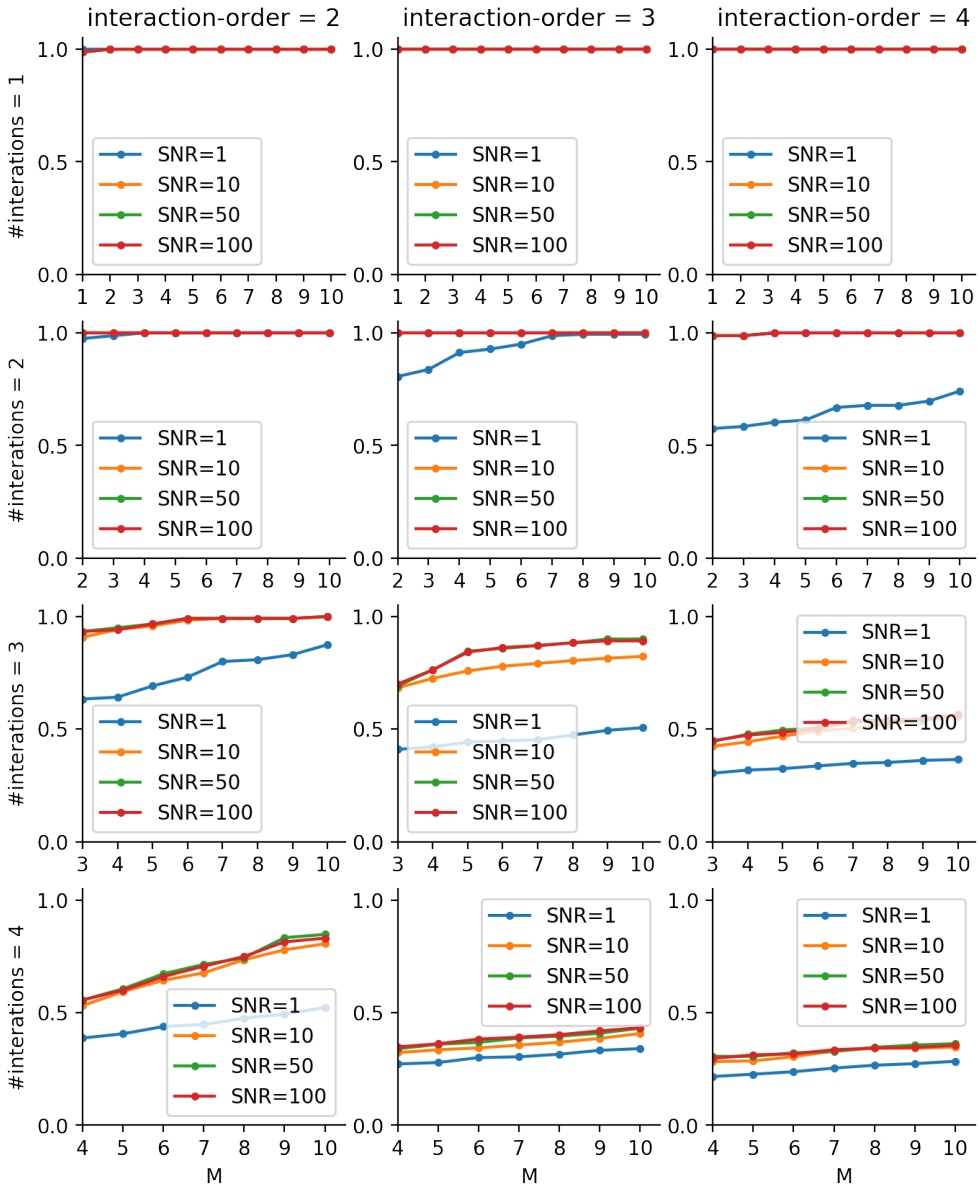


Figure 10.1: Simulation results for different number of basic interactions (#interactions), interaction-orders, and SNRs. The y-axis shows the proximity score in (10.1) against different values of M on the x-axis.

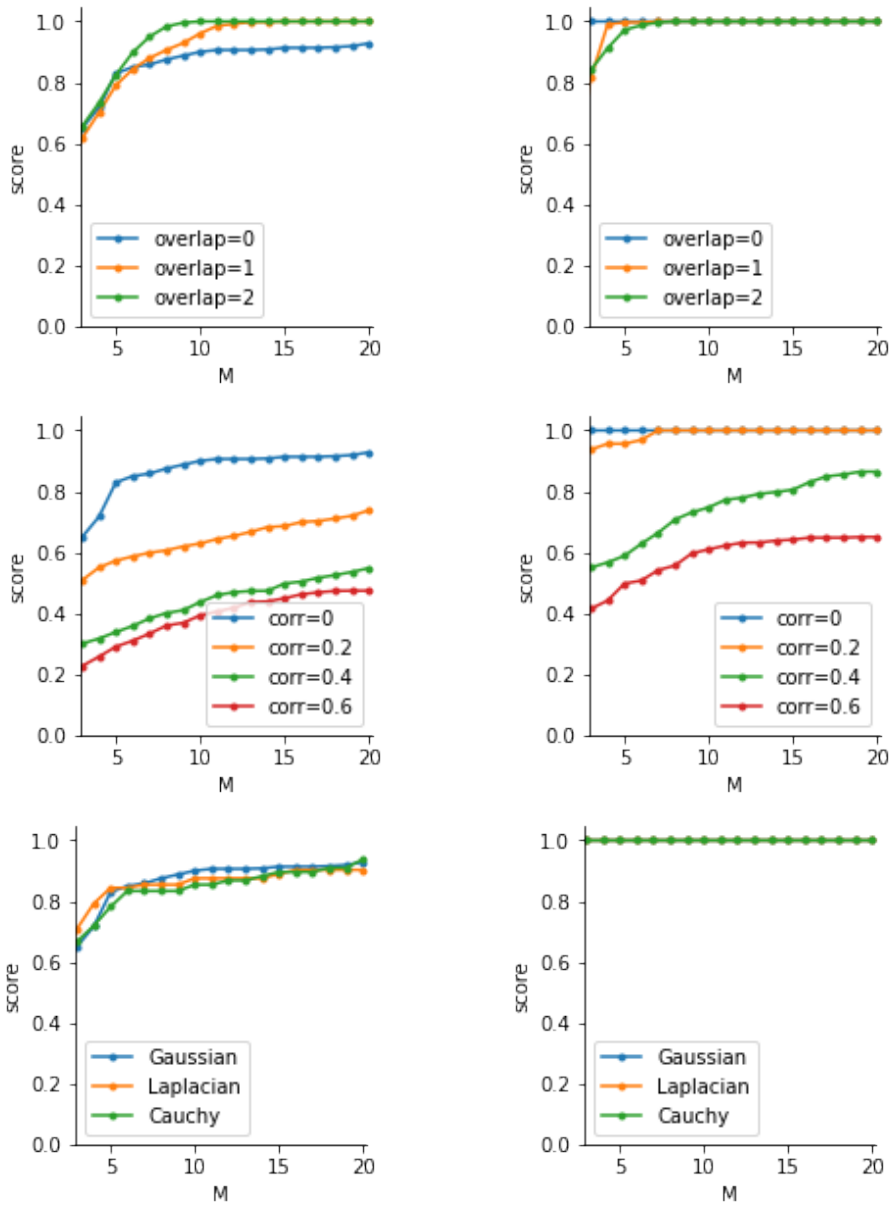


Figure 10.2: Simulation results under LSS model violations: overlapping features (top); correlated features (middle); heavy tailed noise (bottom), see details in the text. Left panel for $K = L = 3$ and right panel for $K = 2$, $L = 3$.

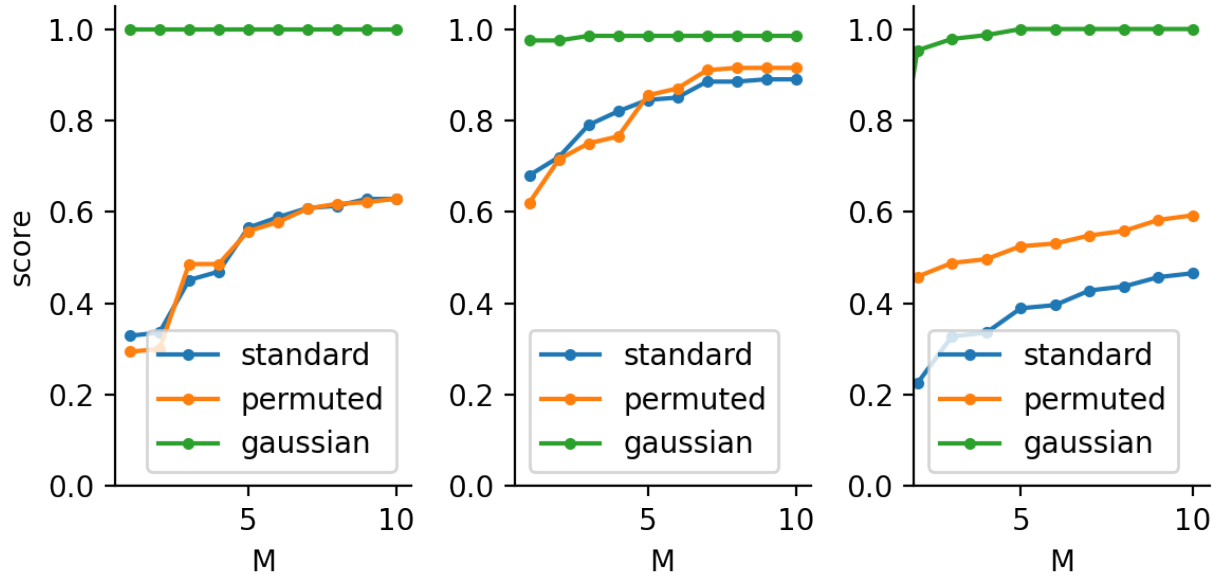


Figure 10.3: Proximity score as in (10.1) plotted against number M of top-M interactions according to LSSrank ranking. Features matrix was taken as in enhancer data from [45] with response as in (AND), (OR), (ADD), see equations (10.2), (10.3), (10.4). "standard" (blue) refers to the case when we use the original feature values of enhancer; "permuted" (orange) refers to the case when we permute each column of the original feature values; "gaussian" (green) refers to the case when features are regenerated via standard Gaussian.

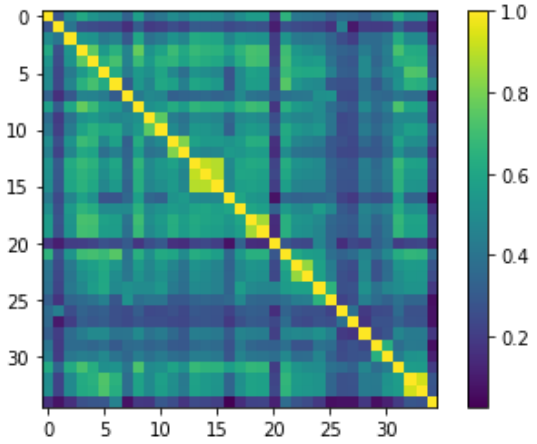


Figure 10.4: Correlation matrix of enhancer features.

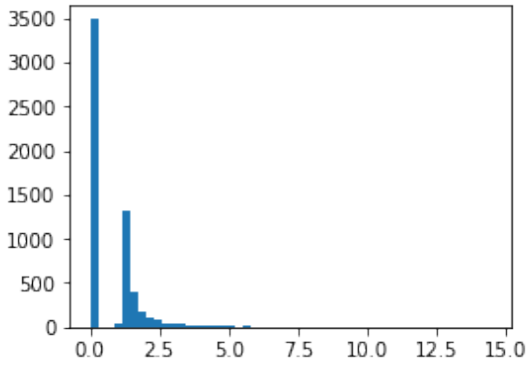


Figure 10.5: Histogram of Kr, one of the features in the enhancer data set.

Chapter 11

Discussion and future work

11.1 Discussion

It is important to study a model that is well scientifically motivated. LSS models provides such a family of models that better reflect certain biological data structures. Also, analyzing ML algorithms under different models can give insights into their empirical adaptivity. Our results are the first to give a theoretical analysis that DWP of an interaction in RF recovers higher order interactions via its decision paths under the LSS model. Moreover, the universality of interaction's DWP in LSS models, gives insights into the general difference between quantitative (e.g. prediction accuracy) and qualitative (e.g. interaction recovery) information extraction. In scientific problems often the latter is of higher interest. Thus, this work narrows the gap between theory and practice and is therefore of general interest to the community.

Our theoretical analysis also gives some insights of RF for practical improvements: In particular, an optimal choice of m_{try} can be extracted from our upper bound, namely, $m_{\text{try}} = (p-2)/(2-s/p)$. For $p \gg s$ this recovers one of the default choices in standard RF implementations, namely, $m_{\text{try}} \approx p/2$, which suggests that with the presence of many noisy features, m_{try}/p should be relatively large. Moreover, our results on impurity decrease thresholding for the set $\hat{\mathcal{F}}_\epsilon$, give theoretical indication why iterative re-weighting in iRF is helpful (see Remark 5).

11.2 Future work

There are a few future works that we plan to do. First, LSSrank gives a possible way to use prevalence to rank interactions. One question is that whether LSSrank can help empirically when incorporated into the current iRF pipeline. Second, for LSSrank it would be helpful to develop algorithms (e.g., similar to FP-growth) that can filter interactions by prevalence depending on the size of the interaction. Third, we would like to extend to more

general interactions models beyond boolean interactions. Finally, it is interesting to study higher-structure recovery for other ML algorithms, e.g., DNN.

Appendix A

Proofs of Part I

A.1 Proof of Proposition 3.2.1

Proof. To prove both models satisfy Assumption I, we just need to prove $|||\cdot|||_{\boldsymbol{\alpha}}$ is lower bounded by $|||\cdot|||_F$ in the linear subspace $H^K = \{A \in \mathbb{R}^{K\times K} | A_{k,k} = 0 \forall\ k\}$. If we can prove $|||\cdot|||_{\boldsymbol{\alpha}}$ is a norm on H^K, then we know it is equivalent to the Frobenious norm since H^K is a finite dimensional space.

In order to show that $|||\cdot|||_{\boldsymbol{\alpha}}$ is a norm, we need to prove three properties:

- Sub-additivity: for any $A, B \in H^K$, $|||A+B|||_{\boldsymbol{\alpha}} \leq |||A|||_{\boldsymbol{\alpha}} + |||B|||_{\boldsymbol{\alpha}}$.
- Absolutely homogeneity: for any $A \in H^K$ and $\lambda > 0$, $|||\lambda A|||_{\boldsymbol{\alpha}} = \lambda |||A|||_{\boldsymbol{\alpha}}$.
- Positive definiteness: If $|||A|||_{\boldsymbol{\alpha}} = 0$ and $A \in H^K$, we know $A = 0$.

The first two properties are quite straightforward so we leave the details to readers. Here we focus on proving the third property. Note that $|||A|||_{\boldsymbol{\alpha}}$ is a sum of K non-negative terms, if $|||A|||_{\boldsymbol{\alpha}} = 0$, then for any $k \in \{1, \ldots, K\}$, each term should be zero, i.e. $\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0$. If $\boldsymbol{\alpha}$ is from Bernoulli-type models $\mathcal{B}(p_1, \ldots, p_K; f)$, then we could further decompose $\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0$ into:

$$\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0 \Leftrightarrow P(\eta_k = 0)\mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0 \Leftrightarrow \mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0.$$

The second "$\Leftrightarrow$" is because $P(\eta_k = 0) = 1 - p_k > 0$ and $A_{k,k} = 0$. Since $\mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0 > P(\eta_1 = \ldots = \eta_K = 1)\mathbb{E}|\sum_j A_{k,j} z_j| \geq 0$ for $p_1, \ldots, p_K \neq 0$, we know $\mathbb{E}|\sum_j A_{k,j} z_j| = 0$. Define $X = \sum_j A_{k,j} z_j$, since $\mathbb{E}|X| = 0$, we know $X = 0$ almost surely. If $A_{j,k}$ are not all zeros, this means $z_1, \ldots, z_K$ are linearly dependent. In other words, z lies in a linear subspace of $\mathbb{R}^K$ almost surely. However, that contradicts the fact that z has a density probability function in $\mathbb{R}^K$. So A must be zero. That completes the proof for Bernoulli-type models. For exact sparse models, the approach is essentially the same.

Now for sparse Gaussian and Bernoulli Gaussian distributions, we can obtained the constant $c_{\boldsymbol{\alpha}}$. We first derive the constant for the sparse Gaussian distribution. For $X \in H^K$,

$$\begin{aligned}
|||X|||_{\boldsymbol{\alpha}} =& \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K}{s}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ k\notin S, |S|=s}} \sqrt{\sum_{j\in S} X_{k,j}^2} \\
=& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K-2}{s-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ k\notin S, |S|=s}} \sqrt{\sum_{j\in S} X_{k,j}^2} \\
\text{(Lemma 6.5 in [100])} \geq& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{K} X_{k,j}^2} \\
(\|x\|_1 \geq \|x\|_2) \quad \geq& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} |||X|||_F.
\end{aligned}$$

Here, we need to use Lemma 6.5 in [100]. For the completeness of this thesis, we rewrite the lemma below:

Lemma 6.5 in [100] Let $z \in \mathbb{R}^{K-1}$, then for $1 \leq l \leq m \leq K-1$,

$$\binom{K-2}{l-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K-1\}\\ |S|=l}} \sqrt{\sum_{j\in S} z_j^2} \geq \binom{K-2}{m-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K-1\}\\ |S|=m}} \sqrt{\sum_{j\in S} z_j^2}.$$

Then the first inequality holds by setting $l = s$ and $m = K-1$. In summary, we have shown that for $X \in H^K$, $|||X|||_{\boldsymbol{\alpha}} \geq \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} |||X|||_F$, which means $c_{\boldsymbol{\alpha}}$ is at least $\frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}}$.

Now, we will compute the constant $c_{\boldsymbol{\alpha}}$ for Bernoulli Gaussian distribution. For $X \in H^K$, if we define $\tilde{s} = \lceil (K-2)p + 1 \rceil$,

$$\begin{aligned}
|||X|||_{\boldsymbol{\alpha}} =& \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \sum_{s=0}^{K-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ |S|=s, k\notin S}} p^s (1-p)^{K-s} \sqrt{\sum_{j\in S} X_{k,j}^2} \\
\text{Lemma 6.6 in [100]} \geq& (1-p) \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K-1}{\tilde{s}}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ k\notin S, |S|=\tilde{s}}} \sqrt{\sum_{j\in S} X_{k,j}^2} \\
\text{Lemma 6.5 in [100]} \geq& (1-p) \frac{\lceil (K-2)p + 1 \rceil}{K-1} \sqrt{\frac{2}{\pi}} |||X|||_F \\
\geq& p(1-p) \sqrt{\frac{2}{\pi}} |||X|||_F.
\end{aligned}$$

Here, we have used Lemma 6.6 in [100]. We rewrite that Lemma using the notations in our thesis as follows:

Lemma 6.6 in [100] Let $p \in (0,1)$ and $\tilde{s} = \lceil (K-2)p + 1 \rceil$. For any $z \in \mathbb{R}^{K-1}$,

$$\sum_{s=0}^{K-1} \sum_{\substack{S \subset \{1,\ldots,K-1\} \\ |S|=s}} p^s (1-p)^{K-1-s} \sqrt{\sum_{j \in S} z_j^2} \geq \binom{K-1}{\tilde{s}}^{-1} \sum_{\substack{S \subset \{1,\ldots,K-1\} \\ |S|=\tilde{s}}} \sqrt{\sum_{j \in S} z_j^2}.$$

In summary, we have shown that for $X \in H^K$, $|||X|||_{\boldsymbol{\alpha}} \geq p(1-p)\sqrt{\frac{2}{\pi}}|||X|||_F$, which means $c_{\boldsymbol{\alpha}}$ is at least $p(1-p)\sqrt{\frac{2}{\pi}}$. □

A.2 Proof of Proposition 3.2.2

Proof. In order to prove Assumption II, we only need to show that for any $c_1, \ldots, c_K$, $P(\sum_{l=1}^d c_l \boldsymbol{\alpha}_l = 0, \text{ and } \exists\ l,\ c_l \boldsymbol{\alpha}_l \neq 0) = 0$. Note that $\boldsymbol{\alpha}_j = \xi_j z_j$ for $j = 1, \ldots, K$,

$$\begin{aligned} & P(\sum_{l=1}^{d} c_l \boldsymbol{\alpha}_l = 0, \text{ and } \exists\ l,\ c_l \boldsymbol{\alpha}_l \neq 0) \\ \leq & \sum_{S \subset \{1\ldots,K\}} P(\xi_l = 1 \text{ if } l \in S \text{ and } 0 \text{ if } l \notin S) \cdot P(\sum_{l \in S} c_l z_l = 0, \text{ and } \sum_{l \in S} c_l^2 > 0). \end{aligned}$$

The inequality holds because $\boldsymbol{\alpha}_k = \eta_k \cdot z_k$ for $k = 1, \ldots, K$ and η and z are independent for exact sparse models or Bernoulli-type models. Since z has a density function, $z_1, \ldots, z_K$ are linearly independent, i.e., $P(\sum_{l \in S} c_l z_l = 0, \text{ and } \sum_{l \in S} c_l^2 > 0) = 0$ for any S. □

A.3 Proofs of Corollaries 4.1.1-4.1.4

Before proving the corollaries, we need the following lemma.

Lemma A.3.1. *If X equals to $c \cdot \mathbf{1}\mathbf{1}^T$, and $|||A|||_{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \sqrt{\frac{2}{\pi}} \frac{s}{K} \binom{K-1}{s-1}^{-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ k \notin S, |S|=s}} \sqrt{\sum_{j \in S} A_{k,j}^2}$, then $|||X|||_{\boldsymbol{\alpha}}^* = \frac{cK(K-1)}{\sqrt{s}(K-s)}\sqrt{\frac{\pi}{2}}$.*

Proof of Lemma A.3.1. Essentially, we are trying to prove that

$$\begin{aligned} \max_{A \neq 0, A \in H^K} \frac{\operatorname{tr}(A^T X)}{|||A|||_{\boldsymbol{\alpha}}} = & \max_{A \neq 0, A \in H^K} \frac{c \sum_{k=1}^K \sum_{j \neq k} A_{k,j}}{\sum_{k=1}^K \frac{s}{K} \binom{K-1}{s-1}^{-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ k \notin S, |S|=s}} \sqrt{\sum_{j \in S} A_{k,j}^2}} \sqrt{\frac{\pi}{2}} \\ = & \frac{cK(K-1)}{\sqrt{s}(K-s)} \sqrt{\frac{\pi}{2}}. \end{aligned}$$

Note that this is equivalent to the fact that the following convex optimization problem attains the minimum $(K-s)\sqrt{s}$:

$$\min \sum_{k=1}^{K} \frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ k\notin S,|S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}$$
$$\text{subject to } \sum_{k=1}^{K}\sum_{j\neq k} A_{k,j} = K(K-1).$$

First of all, note that the problem can be split into K sub-problems: For $k=1,\dots,K$,

$$\min \frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\ k\notin S,|S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}$$
$$\text{subject to } \sum_{j\neq k} A_{k,j} = K-1.$$

Furthermore, note that both the objective and the constraint are permutation symmetric: if $\tilde{A}$ is obtained by permuting off-diagonal elements from each row in A, then the objective function remains the same. It is not hard to show for the optimal solution A^* must satisfy that for any k, $j_1\neq k$, and $j_2\neq k$, $A^*_{k,j_1}=A^*_{k,j_2}$. Therefore, $A^*_{k,j}=1$ and the objective function is $s\binom{K-1}{s-1}^{-1}\binom{K-1}{s}\sqrt{s}=(K-s)\sqrt{s}$. That completes the proof. □

Proof of Corollary 4.1.1. (local identifiability for constant collinearity reference dictionary and sparse Gaussian coefficients) The coefficients are generated from sparse Gaussian distribution $SG(s)$. First, the collinearity matrix $M^*=(\boldsymbol{D}^*)^T\boldsymbol{D}^*=(1-\mu)\mathbb{I}+\mu\mathbf{1}\mathbf{1}^T$. Because $\boldsymbol{\alpha}$ is sparse Gaussian, we know $\mathbb{E}\boldsymbol{\alpha}_j\text{sign}(\boldsymbol{\alpha}_k)=0$ for any $j\neq k$ and $\mathbb{E}|\boldsymbol{\alpha}_j|=\sqrt{\frac{2}{\pi}}\frac{s}{K}$. The bias matrix B is

$$(B(\boldsymbol{\alpha},M^*))_{k,j}=\begin{cases} -M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| = -M_{j,k}\sqrt{\frac{2}{\pi}}\frac{s}{K} = -\sqrt{\frac{2}{\pi}}\frac{\mu s}{K} & \text{for } j\neq k \\ \mathbb{E}|\boldsymbol{\alpha}_j|-\mathbb{E}|\boldsymbol{\alpha}_j|=0 & \text{if } j=k \end{cases}.$$

That means $B(\boldsymbol{\alpha},M^*)$ is a constant matrix except for the diagonal elements. In the proof of Proposition 3.2.1, we showed $|||X|||_{\boldsymbol{\alpha}}=\sqrt{\frac{2}{\pi}}\sum_{k=1}^{K}\frac{s}{K}\binom{K-1}{s-1}^{-1}\sum_{k\notin S,|S|=s}\sqrt{\sum_{j\in S}X_{k,j}^2}$. In general, $|||\cdot|||_{\boldsymbol{\alpha}}^*$ does not have an explicit formula, but for constant matrices, there is a closed form formula (see Lemma A.3.1). Using Lemma A.3.1, we know

$$|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*=\frac{\mu\sqrt{s}(K-1)}{K-s}.$$

Now we will calculate the sharpness constant and the region bound. First of all, $|||\boldsymbol{D}^*|||_2^2=|||M^*|||_2=1+\mu(K-1)$. Second, $|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*=\frac{\mu\sqrt{s}(K-1)}{K-s}$ and $c_{\boldsymbol{\alpha}}\geq\frac{s(K-s)}{K(K-1)}\sqrt{\frac{2}{\pi}}$. Combining

those formulas, the sharpness is at least:

$$\frac{1}{\sqrt{\pi}(1+\mu(K-1))}\frac{s}{K}\left(\frac{K-s}{K-1}-\mu\sqrt{s}\right)\approx\frac{s}{\sqrt{\pi}\mu K^2}\left(1-\mu\sqrt{s}\right)\text{ for large } K.$$

For sparse Gaussian distributions, $\max_j \mathbb{E}|\boldsymbol{\alpha}_j| = \sqrt{\frac{2}{\pi}}\frac{s}{K}$, the set S in Theorem 4.1.2 is

$$\begin{aligned}S&=\left\{\boldsymbol{D}\in\mathbb{B}(\mathbb{R}^K)\middle|\,|||\boldsymbol{D}|||_2\le 2|||\boldsymbol{D}^*|||_2,|||\boldsymbol{D}-\boldsymbol{D}^*|||_F\le\frac{(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)\cdot c_{\boldsymbol{\alpha}}}{8\sqrt{2}|||\boldsymbol{D}^*|||_2^2\max_j\mathbb{E}|\boldsymbol{\alpha}_j|}\right\}\\&=\left\{\boldsymbol{D}\in\mathbb{B}(\mathbb{R}^K)\mid|||\boldsymbol{D}|||_2\le 2\sqrt{1+\mu(K-1)},\right.\\&\qquad\left.|||\boldsymbol{D}-\boldsymbol{D}^*|||_F\le\frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(\frac{K-s}{K-1}-\mu\sqrt{s}\right)\right\},\end{aligned}$$

which completes the proof. □

Proof of Corollary 4.1.2. Assume the reference dictionary is a constant collinearity dictionary with coherence μ and the coefficients are generated from non-negative sparse Gaussian distribution $|SG(s)|$. Since $\mathbb{E}\boldsymbol{\alpha}_k\text{sign}(\boldsymbol{\alpha}_j)=\mathbb{E}\eta_k\eta_j z_k=\sqrt{\frac{2}{\pi}}\frac{s(s-1)}{K(K-1)}$ when $j\neq k$, it can be shown that

$$(B(\boldsymbol{\alpha},M^*))_{k,j}=\begin{cases}-\sqrt{\frac{2}{\pi}}\left(\frac{\mu s}{K}-\frac{s(s-1)}{K(K-1)}\right) & \text{for } j\neq k.\\ 0 & \text{if } j=k.\end{cases}$$

This shows $B(\boldsymbol{\alpha},M^*)$ is still a constant matrix except the diagonal elements. However, compared with standard sparse Gaussian coefficients, the constant here is $\sqrt{\frac{2}{\pi}}\left(\frac{\mu s}{K}-\frac{s(s-1)}{K(K-1)}\right)$, which is smaller than $\sqrt{\frac{2}{\pi}}\frac{\mu s}{K}$ in Corollary 4.1.1. Recall the explanation of the matrix B after Theorem 4.1.1, that is because for non-negative sparse Gaussian coefficients, the bias matrix B_1 introduced by the coefficient is of different signs compared to the bias matrix B_2 introduced by the reference dictionary and they cancel with each other. In standard sparse Gaussian case, $B=0$ if $\mu=0$, which means the reference dictionary is orthogonal. For this non-negative case, $B=0$ if $\mu=s/K$, which means the atoms in the reference dictionary should have positive collinearity s/K. As will be shown next, this significantly relaxes the local identifiability condition for non-negative coefficients.

We now compute the closed form formula for the dual semi-norm. By definition, for any matrix X whose elements are all non-negative, $|||X|||_{\boldsymbol{\alpha}}=\sum_{k=1}^K\mathbb{E}|\sum_{j=1}^K X_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k=0)=\sum_{k=1}^K\sum_{j=1}^K X_{k,j}\mathbb{E}\boldsymbol{\alpha}_j\mathbf{1}(\boldsymbol{\alpha}_k=0)=\sqrt{\frac{2}{\pi}}\frac{s(K-s)}{K(K-1)}\sum_{k=1}^K\sum_{j=1,j\neq k}^K X_{j,k}$. Thus we have

$$|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*=\frac{\sqrt{\frac{2}{\pi}}\frac{s}{K}\cdot\left|\mu-\frac{s-1}{K-1}\right|}{\sqrt{\frac{2}{\pi}}\frac{s(K-s)}{K(K-1)}}=\frac{K-1}{K-s}\cdot\left|\mu-\frac{s-1}{K-1}\right|.$$

□

Proof of Corollary 4.1.3. First of all,

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| = -M_{j,k}\sqrt{\frac{2}{\pi}}p = -\sqrt{\frac{2}{\pi}}\mu p & \text{for } j \neq k. \\ 0 & \text{if } j = k. \end{cases}$$

Because all the elements in the matrix are constant except the diagonal ones, similar to Lemma A.3.1, we can show the optimal A that attains the maximum of $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \max \frac{\mathrm{tr}(A^T B(\boldsymbol{\alpha}, M^*))}{|||A|||_{\boldsymbol{\alpha}}}$ is a constant matrix $\mathbf{1}\mathbf{1}^T$. Thus, we have

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \frac{\mu p(K-1)}{(1-p)\sum_{s=0}^{K-1}\binom{K-1}{s}p^s(1-p)^{K-1-s}\sqrt{s}} \leq \frac{\mu\sqrt{p(K-1)}}{1-p}.$$

Here we are using the Jensen inequality that

$$\sum_{s=0}^{K-1}\binom{K-1}{s}p^s(1-p)^{K-1-s}\sqrt{s} > \sqrt{\sum_{s=0}^{K-1}\binom{K-1}{s}p^s(1-p)^{K-1-s}s} = \sqrt{(K-1)p}.$$

Thus RHS < 1 when μ and p are small. The sharpness is at least

$$\frac{p}{\sqrt{\pi}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right),$$

Because $\mathbb{E}|\boldsymbol{\alpha}_j| = p\sqrt{\frac{2}{\pi}}$ for any j, the set S in Theorem 4.1.2 is

$$\left\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)\middle| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, \right. \\ \left. |||\boldsymbol{D}-\boldsymbol{D}^*|||_F^2 \leq \frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right)\right\}.$$

□

Proof of Corollary 4.1.4. We compute the local identifiability condition when the reference dictionary is a constant collinearity dictionary with coherence μ and the coefficients are generated from sparse Laplace distribution, i.e., for any j $\boldsymbol{\alpha}_j = \xi_j z_j$ where z_j is from a standard Laplace distribution and ξ is a random 0-1 vector with s nonzeros. For standard Laplace distributions, since $\mathbb{E}|\boldsymbol{\alpha}_j| = \frac{s}{K}$, we have

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -\mu\frac{s}{K} & \text{for } j \neq k. \\ 0 & \text{if } j = k. \end{cases}$$

Similar to Lemma A.3.1, we can show the optimal A that attains the maximum of $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*$ is a constant matrix $\mathbf{1}\mathbf{1}^T$.

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \frac{\mu s(K-1)}{(K-s)\iint_0^\infty |y-x|(xy)^{s-1}\exp(-(x+y))\Gamma(s)^{-2}dxdy}.$$

To derive this denominator, we need to give an explicit formula for a linear combination of Laplace random variables. The formula can be found in a few papers, e.g., [63]. □

A.4 Proofs of theorems 4.1.1-4.2.2

The following lemmas are useful for proving Theorem 4.1.1.

Lemma A.4.1. *Given two dictionaries $\boldsymbol{D}$ and $\boldsymbol{D}' \in \mathbb{B}(\mathbb{R}^K)$, we have the decomposition:*

$$\boldsymbol{D}^{-1}\boldsymbol{D}' = \mathbb{I} + (\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')) + \Lambda(\boldsymbol{D}, \boldsymbol{D}').$$

where $\Lambda(\boldsymbol{D}, \boldsymbol{D}')$ is a diagonal matrix whose j-th element is $-\frac{1}{2}\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2$. Then we know

1. *For any $j = 1, \ldots, K$, $M[j,](\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}')) = 0$ where $M = \boldsymbol{D}^T\boldsymbol{D}$.*
2. *$|||\Lambda(\boldsymbol{D})|||_F = \Theta(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2)$:*

$$\frac{1}{2\sqrt{K}}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \leq |||\Lambda(\boldsymbol{D})|||_F \leq \frac{1}{2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2.$$

3. *When $\langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle \geq 0$ for any $j = 1, \ldots, K$, $|||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F = \Theta(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F)$:*

$$\frac{|||\boldsymbol{D} - \boldsymbol{D}'|||_F}{\sqrt{2}|||\boldsymbol{D}|||_2} \leq |||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F \leq |||\boldsymbol{D}^{-1}|||_2 \cdot |||\boldsymbol{D} - \boldsymbol{D}'|||_F$$

4. *Let $M' = (\boldsymbol{D}')^T\boldsymbol{D}'$, for any A satisfying $M'[j,]A_j = 0$ for $j = 1, \ldots, K$ and $|||A|||_F$ sufficiently small, there is a $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)$ such that $\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}') = A$.*

Proof of Lemma A.4.1. **(1)**:

$$\begin{aligned}
& M[j,](\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}')) && \text{(A.1)}\\
=& \langle \boldsymbol{D}_j, \boldsymbol{D}(\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}')) \rangle && \text{(A.2)}\\
=& \langle \boldsymbol{D}_j, \boldsymbol{D}'_j - \boldsymbol{D}_j + \frac{1}{2}\boldsymbol{D}_j\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2 \rangle && \text{(A.3)}\\
=& \langle \boldsymbol{D}_j, \boldsymbol{D}'_j - \boldsymbol{D}_j \rangle + \frac{1}{2}\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2 && \text{(A.4)}\\
=& \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle - 1 + 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle = 0. && \text{(A.5)}
\end{aligned}$$

(2): $|||\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)|||_F = \frac{1}{2}\sqrt{\sum_j \|\boldsymbol{D}_j - \boldsymbol{D}_j^*\|_2^4} \leq \frac{1}{2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2$. On the other hand, because of the power inequality $\|x\|_2 \geq \frac{1}{\sqrt{K}}\|x\|_1$, we have

$$|||\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)|||_F = \frac{1}{2}\sqrt{\sum_j \|\boldsymbol{D}_j - \boldsymbol{D}_j^*\|_2^4} \geq \frac{1}{2\sqrt{K}}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2.$$

(3): First, consider $|||\boldsymbol{D}' - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F$, we have

$$\begin{aligned}|||\boldsymbol{D}' - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F^2 &= \sum_{j=1}^{K} \|\boldsymbol{D}'_j - \boldsymbol{D}_j\langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle\|_2^2 \\ &= \sum_{j=1}^{K} 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle^2 \\ &= \sum_{j=1}^{K} \min_{t_j \in \mathbb{R}} \|\boldsymbol{D}'_j - t_j \cdot \boldsymbol{D}_j\|_2^2.\end{aligned}$$

Then by taking $t_j = 1$ for all $j = 1, \ldots, K$, we have

$$|||\boldsymbol{D}' - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F^2 \leq |||\boldsymbol{D}' - \boldsymbol{D}|||_F^2.$$

On the other hand, when $\langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle \geq 0$.

$$\sum_{j=1}^{K} 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle^2 = \sum_{j=1}^{K} (1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle)(1 + \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle) \geq \sum_{j=1}^{K} (1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle) = \frac{1}{2}|||\boldsymbol{D} - \boldsymbol{D}'|||_F^2.$$

Then for $(\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}'))$, using the above inequalities, we have:

$$\begin{aligned}|||\boldsymbol{D}' - \boldsymbol{D}|||_F \leq& \sqrt{2}|||\boldsymbol{D}' - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F && \text{(A.6)} \\ \leq& \sqrt{2}|||\boldsymbol{D}|||_2|||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F, && \text{(A.7)}\end{aligned}$$

which proves the first inequality. The second inequality follows from

$$\begin{aligned}&|||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F && \text{(A.8)} \\ \leq& |||\boldsymbol{D}^{-1}|||_2|||\boldsymbol{D}' - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F && \text{(A.9)} \\ \leq& |||\boldsymbol{D}^{-1}|||_2|||\boldsymbol{D}' - \boldsymbol{D}|||_F. && \text{(A.10)}\end{aligned}$$

(4): Consider a differentiable mapping $F(\boldsymbol{D}) = \boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')$ from $\mathbb{B}(\mathbb{R}^K)$ to a linear manifold

$$H = \{A \in \mathbb{R}^{K\times K} \Big| M'[j,]A_j = 0 \text{ for any } j = 1, \ldots, K.\}$$

Since $F(\boldsymbol{D}') = \boldsymbol{0}$, if we can prove the differential of F at $\boldsymbol{D}'$, namely dF, is bijective from the tangent space $T\mathbb{B}(\mathbb{R}^K)\Big|_{\boldsymbol{D}'} = \{A \in \mathbb{R}^{K\times K} \Big| \langle D'_j, A_j\rangle = 0 \text{ for any } j = 1, \ldots, K.\}$ to the tangent space $TH\Big|_0 = H$, then by the inverse function theorem on the manifold, we have the conclusion. To prove it is indeed bijective, we note that $dF(\Delta)\Big|_{\boldsymbol{D}'}$ is $\sum_{k,j}(\boldsymbol{D}')_j^{-1}\mathbb{I}[k,]\Delta_{j,k} = (\boldsymbol{D}')^{-1}\Delta$.

Clearly dF is injective: $(\boldsymbol{D}')^{-1}\Delta = 0$ implies $\Delta = 0$. To show it is also surjective, first of all for any $\Delta \in T\mathbb{B}(\mathbb{R}^K)\Big|_{\boldsymbol{D}'}$, its image under dF is in H:

$$\begin{aligned} M'[j,](\boldsymbol{D}')^{-1}\Delta_j =& \left\langle \boldsymbol{D}'_j, \boldsymbol{D}'((\boldsymbol{D}')^{-1}\Delta_j \right\rangle \\ =& \left\langle \boldsymbol{D}'_j, \Delta_j \right\rangle = 0. \end{aligned}$$

Because these two linear manifolds have the same dimension, dF must be one-on-one. This concludes the proof. □

Lemma A.4.2. *If $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular with constant $c_{\boldsymbol{\alpha}}$, then we know for any $\boldsymbol{D}, \boldsymbol{D}'$ such that $\left\langle \boldsymbol{D}_j, \boldsymbol{D}'_j \right\rangle \geq 0$ for any $j = 1, \ldots, K$, $|||(\boldsymbol{D})^{-1}\boldsymbol{D}'|||_{\boldsymbol{\alpha}} \geq \frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}'|||_F$.*

Proof. First of all, because for any $A \in \mathbb{R}^{K\times K}$, by definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, $|||A|||_{\boldsymbol{\alpha}}$ does not depend on diagonal elements $A_{j,j}$ for any $j = 1, \ldots, K$. Thus, $|||(\boldsymbol{D})^{-1}\boldsymbol{D}'|||_{\boldsymbol{\alpha}} = |||(\boldsymbol{D})^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_{\boldsymbol{\alpha}}$, where Λ is defined in Lemma A.4.1. If we denote A as $(\boldsymbol{D})^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')$, then Lemma A.4.1 shows $M[j,]A_j = 0$. Since $M_{j,j} = 1$, $A_{j,j} = -M[j,-j]A[-j,j]$. Thus

$$\|A_j\|_2^2 \leq (M[j,-j]A[-j,j])^2 + \|A[-j,j]\|_2^2 \leq (\|M[j,-j]\|_2^2 + 1)\|A[-j,j]\|_2^2 = \|M[j,]\|_2^2\|A[-j,j]\|_2^2.$$

Summing over j, we have

$$|||A|||_F \leq \max_j \|M[j,]\|_2 \sqrt{\sum_j \|A[-j,j]\|_2^2}.$$

Note that for any j, $\|M[j,]\|_2 = \|\boldsymbol{D}_j^T\boldsymbol{D}\|_2 \leq |||\boldsymbol{D}|||_2$, thus we have:

$$|||A|||_F \leq |||\boldsymbol{D}|||_2 \sqrt{\sum_j \|A[-j,j]\|_2^2}.$$

On the other hand, by Lemma A.4.1, we know $|||A|||_F \geq \frac{1}{\sqrt{2}|||\boldsymbol{D}|||_2}|||\boldsymbol{D} - \boldsymbol{D}'|||_F$. Combining those together, we have

$$\begin{aligned} |||(\boldsymbol{D})^{-1}\boldsymbol{D}'|||_{\boldsymbol{\alpha}} =& |||(\boldsymbol{D})^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_{\boldsymbol{\alpha}} \\ \text{(Because of Assumption I))} \geq& c_{\boldsymbol{\alpha}} \sqrt{\sum_j \|A[-j,j]\|_2^2} \\ \geq& c_{\boldsymbol{\alpha}} \frac{|||A|||_F}{|||\boldsymbol{D}|||_2} \\ \geq& c_{\boldsymbol{\alpha}} \frac{|||\boldsymbol{D} - \boldsymbol{D}'|||_F}{\sqrt{2}|||\boldsymbol{D}|||_2^2}. \end{aligned}$$

□

Lemma A.4.3. *For $x, y \in \mathbb{R}$, $y \cdot \text{sign}(x) + |x| + |y| \cdot \mathbf{1}(x = 0) \leq |y + x| \leq y \cdot \text{sign}(x) + |x| + |y| \cdot \mathbf{1}(x = 0) + 2|y| \cdot \mathbf{1}(|y| > |x| > 0)$.*

Proof. If $x = 0$, the above inequality definitely holds. So without loss of generality, let's assume $x \neq 0$. When $|y| < |x|$, $\text{sign}(x + y) = \text{sign}(x)$, so $|y + x| = \text{sign}(x)(x + y) = |x| + \text{sign}(x)y$. When $|y| > |x|$, $\text{sign}(x + y) = \text{sign}(y)$, if $\text{sign}(x) = \text{sign}(y)$, clearly we have $|x + y| = |x| + |y| = y\text{sign}(x) + |x|$. If $\text{sign}(x) \neq \text{sign}(y)$, $|y + x| = |y| - |x| > |x| - |y| = |x| + y\text{sign}(x)$. So in summary, we prove the first inequality. The second inequality comes from: $|y + x| \leq |y| + |x| \leq |x| + 2|y| + y\text{sign}(x)$, which completes the proof. □

Lemma A.4.4. *We have the upper and lower bound of the objective function:*

$$\begin{aligned}
&\mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \text{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) + o(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F) \\
&\geq \mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \\
&\mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \text{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1
\end{aligned}$$

Proof of Lemma A.4.4. By Lemma A.4.1, $(\boldsymbol{D})^{-1}\boldsymbol{D}^*$ can be decomposed into

$$\boldsymbol{D}^{-1}\boldsymbol{D}^* = \mathbb{I} + \Delta(\boldsymbol{D}, \boldsymbol{D}^*) + \Lambda(\boldsymbol{D}, \boldsymbol{D}^*),$$

where $\Delta(\boldsymbol{D}, \boldsymbol{D}^*) = \boldsymbol{D}^{-1}\boldsymbol{D}^* - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ and $\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ is defined in Lemma A.4.1. In what follows, we use Λ, Δ without writing $\boldsymbol{D}, \boldsymbol{D}^*$ explicitly for notation ease.

Let $\Delta_{k,j}$ be the element of Δ at k-th row and j-th column. Then the objective function can be lower bounded by:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 =& \mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x} - (\mathbb{I} - \boldsymbol{D}^{-1}\boldsymbol{D}^*)(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 && \text{(A.11)} \\
=& \mathbb{E}\|\boldsymbol{\alpha} + (\Delta + \Lambda)\boldsymbol{\alpha}\|_1 && \text{(A.12)} \\
(a)\ \geq& \mathbb{E}\|\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}\|_1 - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 && \text{(A.13)} \\
(b)\ \geq& \mathbb{E}\sum_{k=1}^{K} |\boldsymbol{\alpha}_k| + \mathbf{1}(\boldsymbol{\alpha}_k = 0)|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| - \text{sign}\boldsymbol{\alpha}_k \sum_j \Delta_{k,j}\boldsymbol{\alpha}_j - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 && \text{(A.14)} \\
\geq& \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\Delta|||_{\boldsymbol{\alpha}} - \mathbb{E}\sum_{k,j} \Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j\text{sign}\boldsymbol{\alpha}_k - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1. && \text{(A.15)}
\end{aligned}$$

(a) holds because of the triangle inequality. (b) holds because of Lemma A.4.3 (let $x = \Delta[k,]\boldsymbol{\alpha}$ and $y = \boldsymbol{\alpha}_k$). Note that by the definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, the diagonal elements of Δ do not matter, so $|||\Delta|||_{\boldsymbol{\alpha}} = |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}$.

Recall $M_{j,k} = \langle \boldsymbol{D}_j, \boldsymbol{D}_k \rangle$, by Lemma A.4.1, $\Delta_{k,j}$ satisfies: $M[j,]\Delta_j = \sum_{k\neq j} M_{j,k}\Delta_{k,j} + \Delta_{j,j} = 0$ (Because $M_{j,j} = 1$) for any j. Thus we have

$$\sum_{j,k=1}^{K} \Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j \text{sign}\boldsymbol{\alpha}_k = \sum_{j=1}^{K}\left(\sum_{k\neq j}\Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j\text{sign}\boldsymbol{\alpha}_k + \Delta_{j,j}\mathbb{E}|\boldsymbol{\alpha}_j|\right) \tag{A.16}$$

$$= \sum_{j=1}^{K}\sum_{k\neq j}\Delta_{k,j}\left(\mathbb{E}\boldsymbol{\alpha}_j\text{sign}\boldsymbol{\alpha}_k - M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j|\right) = \text{tr}(B(\boldsymbol{\alpha}, M)^T\Delta). \tag{A.17}$$

Because the diagonal elements of $B(\boldsymbol{\alpha}, M)$ are all zeros, we know

$$\text{tr}(B(\boldsymbol{\alpha}, M)^T\Delta) = \text{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*).$$

In summary, we have shown that

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \text{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1. \tag{A.18}$$

In order to have an upper bound, we have

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 = \mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x} - (\mathbb{I} - \boldsymbol{D}^{-1}\boldsymbol{D}^*)(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 \tag{A.19}$$

$$= \mathbb{E}\|\boldsymbol{\alpha} + (\Delta + \Lambda)\boldsymbol{\alpha}\|_1 \tag{A.20}$$

$$\leq \mathbb{E}\|\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}\|_1 + \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \tag{A.21}$$

$$\text{(Lemma A.4.3)} \leq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \text{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) \tag{A.22}$$

$$+ \sum_k 2\mathbb{E}\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| \mathbf{1}\left(\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| > |\boldsymbol{\alpha}_k| > 0\right) + \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1. \tag{A.23}$$

Note that by Lemma A.4.1, $\mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \leq |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \max_j \mathbb{E}|\boldsymbol{\alpha}_j| = o(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F)$. Furthermore,

$$\mathbb{E}\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| \mathbf{1}\left(\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| > |\boldsymbol{\alpha}_k| > 0\right) \tag{A.24}$$

$$\leq \sum_{k=1}^{K} \max_j |\Delta_{k,j}| \cdot \mathbb{E}\ \mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1. \tag{A.25}$$

Because $\mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 \leq \|\boldsymbol{\alpha}\|_1$, $\mathbb{E}\|\boldsymbol{\alpha}\|_1 < \infty$, and

$$\lim_{\Delta_{k,j}\to 0} \mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 = 0 \quad a.s.,$$

by the dominant convergence theorem, we know

$$\begin{aligned}&\lim_{\Delta\to 0}\mathbb{E}\mathbf{1}(\boldsymbol{\alpha}_k\neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j|\geq|\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1\\ =&\mathbb{E}\lim_{\Delta\to 0}\mathbf{1}(\boldsymbol{\alpha}_k\neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j|\geq|\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1\\ =&0.\end{aligned}$$

This means (A.24) is $o(|||\boldsymbol{D}-\boldsymbol{D}^*|||_F)$, which proves the upper bound. □

Proof of Theorem 4.1.1. **(i)**: We will first prove that if $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular with constant $c_{\boldsymbol{\alpha}}$ and (4.2) holds, $\boldsymbol{D}^*$ is a sharp local minimum. When (4.2) is satisfied and $\boldsymbol{D}\to\boldsymbol{D}^*$, $|||B(\boldsymbol{\alpha},M)|||^*_{\boldsymbol{\alpha}}\to|||B(\boldsymbol{\alpha},M^*)|||^*_{\boldsymbol{\alpha}}<1$ and

$$\begin{aligned}&|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}-\mathrm{tr}(B(\boldsymbol{\alpha},M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*)\\ =&|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}-\mathrm{tr}(B(\boldsymbol{\alpha},M^*)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*)+o(|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}})\\ \geq&(1-|||B(\boldsymbol{\alpha},M)|||^*_{\boldsymbol{\alpha}})|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}+o(|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}).\end{aligned}$$

Because $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular and Lemma A.4.2, by appropriately choosing signs of each column in $\boldsymbol{D}^*$, we have

$$|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}\geq\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D}^*-\boldsymbol{D}|||_F.$$

Combine those two inequalities, when $|||\boldsymbol{D}-\boldsymbol{D}^*|||_F$ is small enough,

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1-\mathbb{E}\|\boldsymbol{\alpha}\|_1\\ \geq&(1-|||B(\boldsymbol{\alpha},M^*)|||^*_{\boldsymbol{\alpha}})\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}\cdot|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D}-\boldsymbol{D}^*|||_F+o(|||\boldsymbol{D}-\boldsymbol{D}^*|||_F).\end{aligned}$$

By Definition 3.2.1, $\boldsymbol{D}^*$ is a sharp local minimum with sharpness at least

$$(1-|||B(\boldsymbol{\alpha},M^*)|||^*_{\boldsymbol{\alpha}})\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}^*|||_2^2}.$$

(ii) When (4.2) does not hold or $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular, $\boldsymbol{D}^*$ is not a sharp local minimum.

If $|||B(\boldsymbol{\alpha},M^*)|||^*_{\boldsymbol{\alpha}}\geq 1$, then there exists Δ' such that $|||\Delta'|||_{\boldsymbol{\alpha}}-\mathrm{tr}(B(\boldsymbol{\alpha},M^*)^T\Delta')\leq 0$. Note that the left hand side does not depend on diagonal elements of Δ', so we can find a matrix Δ that is the same as Δ' except the diagonal elements such that $M^*[j,]\Delta_j=0$ for any j and $|||\Delta|||_{\boldsymbol{\alpha}}-\mathrm{tr}(B(\boldsymbol{\alpha},M^*)^T\Delta)\leq 0$. For any $t>0$, by Lemma A.4.1 we can construct a series of dictionaries $\boldsymbol{D}(t)$ for a sufficiently small t such that

$$(\boldsymbol{D}(t))^{-1}\boldsymbol{D}^*=\mathbb{I}+\Delta+\Lambda(\boldsymbol{D}(t),\boldsymbol{D}^*).$$

Then by Lemma A.4.4, we have the formula for the objective of $\boldsymbol{D}(t)$:

$$\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 = \mathbb{E}\|\boldsymbol{\alpha}\|_1 + \left(|||\Delta|||_{\boldsymbol{\alpha}} - \operatorname{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta)\right) + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F). \tag{A.26}$$

Because $|||\Delta|||_{\boldsymbol{\alpha}} - \operatorname{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta) \leq 0$, $\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 \leq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F)$. By definition, $\boldsymbol{D}^*$ is not a sharp local minimum. If $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular, for any $c > 0$, there exists Δ such that $M^*[j,]\Delta_j = 0$ for any j and $|||\Delta|||_{\boldsymbol{\alpha}} < c\|\Delta\|_F$. Without loss of generality, assume $\operatorname{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta) \geq 0$, otherwise just take $-\Delta$. For sufficiently small t, there exists a dictionary $\boldsymbol{D}(t)$ such that

$$(\boldsymbol{D}(t))^{-1}\boldsymbol{D}^* = \mathbb{I} + \Delta + \Lambda(\boldsymbol{D}(t), \boldsymbol{D}^*).$$

Then by Lemma A.4.4, we have the formula for the objective of $\boldsymbol{D}(t)$:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 =& \mathbb{E}\|\boldsymbol{\alpha}\|_1 + \left(|||\Delta|||_{\boldsymbol{\alpha}} - \operatorname{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta)\right) + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F) && \text{(A.27)}\\
\leq& \mathbb{E}\|\boldsymbol{\alpha}\|_1 + c|||\Delta|||_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F) && \text{(A.28)}\\
\leq& \mathbb{E}\|\boldsymbol{\alpha}\|_1 + c|||\boldsymbol{D}(t)^{-1}|||_2 \cdot |||\boldsymbol{D}(t) - \boldsymbol{D}^*|||_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F) && \text{(A.29)}
\end{aligned}$$

Because that holds for any $c > 0$, by definition, we have shown $\boldsymbol{D}^*$ is not a sharp local minimum.

(iii): When $|||B(\boldsymbol{\alpha}, M^*)|||^*_{\boldsymbol{\alpha}} > 1$, $\boldsymbol{D}^*$ is not a local minimum. This part is essentially the same as (ii). The key is to construct a series of dictionaries $\boldsymbol{D}(t)$ using Lemma A.4.1 as in (ii). Then by using the upper bound in Lemma A.4.4, we can find a small $t > 0$ and a small $c > 0$ such that

$$\mathbb{E}\|\boldsymbol{D}_t^{-1}\boldsymbol{x}\|_1 \leq \mathbb{E}\|\boldsymbol{\alpha}\|_1 - c\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F).$$

Thus by definition $\boldsymbol{D}^*$ is not a local minimum. □

Proof of Theorem 4.1.2. Note that by the definition of $\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ as in Lemma A.4.1, we have

$$\mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1 \leq \max_j \mathbb{E}|\boldsymbol{\alpha}_j| |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \tag{A.30}$$

On the other hand, by Lemma A.4.4, we know

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 - \mathbb{E}\|\boldsymbol{\alpha}\|_1 \geq |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \operatorname{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1$$

Similar to the proof of Theorem 4.1.1, the right hand side is bounded by

$$\begin{aligned}
&|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \operatorname{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1\\
\geq& (1 - |||B(\boldsymbol{\alpha}, M)|||^*_{\boldsymbol{\alpha}})|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1\\
\geq& (1 - |||B(\boldsymbol{\alpha}, M)|||^*_{\boldsymbol{\alpha}})|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2\\
\geq& (1 - |||B(\boldsymbol{\alpha}, M)|||^*_{\boldsymbol{\alpha}})\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 && \text{(A.31)}\\
\geq& (1 - |||B(\boldsymbol{\alpha}, M)|||^*_{\boldsymbol{\alpha}})\frac{c_{\boldsymbol{\alpha}}}{4\sqrt{2}|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2. && \text{(A.32)}
\end{aligned}$$

Because $|||M-M^*|||_F \le (|||\boldsymbol{D}|||_2+|||\boldsymbol{D}^*|||_2)\cdot|||\boldsymbol{D}-\boldsymbol{D}^*|||_F \le 3|||\boldsymbol{D}^*|||_2\cdot|||\boldsymbol{D}-\boldsymbol{D}^*|||_F$ and $|||\boldsymbol{D}-\boldsymbol{D}^*|||_F \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{8\sqrt{2}\max_j \mathbb{E}|\boldsymbol{\alpha}_j|\,|||\boldsymbol{D}^*|||_2^2}$ we know $|||M-M^*|||_F \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{2\max_j \mathbb{E}|\boldsymbol{\alpha}_j|\,|||\boldsymbol{D}^*|||_2} \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{2\max_j \mathbb{E}|\boldsymbol{\alpha}_j|}$. The last inequality is because $|||\boldsymbol{D}^*|||_2 \ge 1$. Based on this chain of inequalities, we have:

$$\begin{aligned}
1-|||B(\boldsymbol{\alpha},M)|||_{\boldsymbol{\alpha}}^* \ge& 1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^* - \big||||B(\boldsymbol{\alpha},M)|||_{\boldsymbol{\alpha}}^* - |||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*\big| \\
\ge& 1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^* - |||B(\boldsymbol{\alpha},M)-B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^* \\
\ge& 1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^* - \frac{1}{c_{\boldsymbol{\alpha}}}|||B(\boldsymbol{\alpha},M)-B(\boldsymbol{\alpha},M^*)|||_F \\
\ge& 1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^* - \frac{1}{c_{\boldsymbol{\alpha}}}\max_j \mathbb{E}|\boldsymbol{\alpha}_j|\cdot|||M-M^*|||_F \\
\ge& \frac{1}{2}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*).
\end{aligned}$$

Based on this, (A.32) is bounded by:

$$\frac{1}{2}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{4\sqrt{2}|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D}-\boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j|\cdot|||\boldsymbol{D}-\boldsymbol{D}^*|||_F^2 \tag{A.33}$$

$$\ge \left(\frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{8\sqrt{2}\max_j \mathbb{E}|\boldsymbol{\alpha}_j|\,|||\boldsymbol{D}^*|||_2^2} - |||\boldsymbol{D}-\boldsymbol{D}^*|||_F\right)|||\boldsymbol{D}-\boldsymbol{D}^*|||_F \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \ge 0. \tag{A.34}$$

This shows the LHS is positive when $\boldsymbol{D}\neq\boldsymbol{D}^*$ and we have completed the proof. □

Proof of Theorem 4.2.1. In order to prove Theorem 4.2.1, it suffices to prove any dictionary $\boldsymbol{D}$ in $\mathbb{B}(\mathbb{R}^K)$ other than $\boldsymbol{D}^*$ will not be a sharp local minimum. Recall $\boldsymbol{\beta}(\boldsymbol{D})$ is the coefficient of the samples under dictionary $\boldsymbol{D}$, i.e., $\boldsymbol{\beta}(\boldsymbol{D})=\boldsymbol{D}^{-1}\boldsymbol{x}$. For notation ease, we omit $\boldsymbol{D}$ and simply write $\boldsymbol{\beta}$.

The following lemma provides a necessary condition for a dictionary to be a sharp local minimum.

Lemma A.4.5. *For any dictionary $\boldsymbol{D}$, if $\boldsymbol{D}$ is a sharp local minimum of optimization form (4.1), then for any $k=1,\dots,K$, $\boldsymbol{\beta}\cdot\mathbf{1}(\boldsymbol{\beta}_k=0)$ does not lie in any linear subspace of dimension $K-2$.*

Proof of Lemma A.4.5. If $\boldsymbol{D}$ is a sharp local minimum, by the proof of Theorem 4.1.1, it should satisfy (A.35).

$$\sum_{j,k}\Delta_{k,j}(\mathbb{E}\boldsymbol{\beta}_j\mathrm{sign}(\boldsymbol{\beta}_k)-M_{j,k}\mathbb{E}|\boldsymbol{\beta}_j|) < \sum_k \mathbb{E}|\sum_j \Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k=0). \tag{A.35}$$

For any $\Delta_{k,j}$, let $\Delta'_{k,j}\triangleq-\Delta_{k,j}$, it should also satisfy (A.35). That makes

$$-\sum_{j,k}\Delta_{k,j}(\mathbb{E}\boldsymbol{\beta}_j\mathrm{sign}(\boldsymbol{\beta}_k)-M_{j,k}\mathbb{E}|\boldsymbol{\beta}_j|) < \sum_k \mathbb{E}|\sum_j \Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k=0).$$

Thus we have

$$\mathbb{E}|\sum_{j=1,j\neq k}^{K}\Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k=0)>0. \tag{A.36}$$

If $\boldsymbol{\beta}\mathbf{1}(\boldsymbol{\beta}_k=0)$ lies in a linear subspace of dimension $K-2$, because there are $K-1$ free parameters in $\Delta_{j,k}$ for $j\neq k$, we can find a set of nonzero $\Delta_{j,k}$ such that $\sum_{j=1,j\neq k}^{K}\Delta_{k,j}\boldsymbol{\beta}_j\cdot\mathbf{1}(\boldsymbol{\beta}_k=0)=0$ a.s.. That contradicts (A.36). Therefore, $\boldsymbol{\beta}\mathbf{1}(\boldsymbol{\beta}_k=0)$ does not lie in any linear subspace of dimension $K-2$. □

In order to show $\boldsymbol{D}\neq\boldsymbol{D}^*$ up to sign-permutation is not a sharp local minimum, by Lemma A.4.5, it suffices to find a k such that the random vector $\boldsymbol{\beta}\cdot\mathbf{1}(\boldsymbol{\beta}_k=0)$ lies in a linear manifold of dimension at most $K-2$.

Note that $\boldsymbol{\beta}=\boldsymbol{D}^{-1}\boldsymbol{D}^*\boldsymbol{\alpha}$ is linear transform of $\boldsymbol{\alpha}$. For $\boldsymbol{D}\neq\boldsymbol{D}^*$ up to the sign-permutation sense, $\boldsymbol{D}^{-1}\boldsymbol{D}^*\neq\mathbb{I}$, which means there exists k such that $\boldsymbol{\beta}_k\neq\boldsymbol{\alpha}_{k'}$ for any $k'=1,\ldots,K$. This means $\boldsymbol{\beta}_k$ is the linear combination of at least two elements in $\boldsymbol{\alpha}$. Without loss of generality, $\boldsymbol{\beta}_k=\sum_{l=1}^{T}c_l\boldsymbol{\alpha}_l$ such that $c_1,\ldots,c_T\neq0$ and $T\geq2$. Because of Assumption II, $\boldsymbol{\beta}_k=0$ implies $\boldsymbol{\alpha}_1=\ldots=\boldsymbol{\alpha}_T=0$. Thus, $\boldsymbol{\beta}\cdot\mathbf{1}(\boldsymbol{\beta}_k=0)=\boldsymbol{D}^{-1}\boldsymbol{D}^*\boldsymbol{\alpha}\mathbf{1}(\boldsymbol{\alpha}_1=\ldots=\boldsymbol{\alpha}_T=0)$, we know $\boldsymbol{\beta}\cdot\mathbf{1}(\boldsymbol{\beta}_k=0)$ lies in a linear manifold of dimension $K-T$ almost surely. □

Proof of Theorem 4.2.2. The whole proof consists of two major steps. The first step is to show that the finite population satisfies the Assumption I with high probability: for any $\epsilon>0$,

$$\begin{aligned}&P\left(\sup_{c_1,\ldots,c_K}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\sum_j c_j\boldsymbol{\alpha}_j^{(i)}=0\text{ and }\sum_{j=1}^{K}\left(c_j\boldsymbol{\alpha}_j^{(i)}\right)^2>0)\geq\epsilon\right)\\ \leq&4\exp\left(2K\left(\ln\frac{n}{2K}+1\right)-n\left(\epsilon-\frac{1}{n}\right)^2\right).\end{aligned} \tag{A.37}$$

In order to prove (A.37), define

$$f_c(\boldsymbol{\alpha})\triangleq\mathbf{1}(\sum_{j=1}^{K}c_j\boldsymbol{\alpha}_j=0\text{ and }\sum_{j=1}^{K}(c_j\boldsymbol{\alpha}_j)^2>0),$$

$\mathcal{F}(\boldsymbol{\alpha})\triangleq\{f_c(\cdot)|c\in\mathbb{R}^K\}$ and consider its VC dimension. We will prove the VC dimension of $\mathcal{F}$ is no bigger than $2K$, namely, for any $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(2K)}$, define a set

$$\mathcal{F}^{(2K)}(\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(2K)})\triangleq\{(f_c(\boldsymbol{\alpha}^{(1)}),\ldots,f_c(\boldsymbol{\alpha}^{(2K)}))|c\in\mathbb{R}^K\},$$

The cardinality of $\mathcal{F}^{(2K)}$ is not 2^{2K}. If $\underbrace{(1,\ldots,1)}_{2K}$ is not in $\mathcal{F}^{(2K)}$, then we are done. Otherwise, there exists c s.t. $f_c(\boldsymbol{\alpha}^{(i)})=1$ for any $i=1,\ldots,2K$. That means $\sum_j c_j\boldsymbol{\alpha}_j^{(i)}=0$ for any $i=1,\ldots,2K$. Therefore, the dimension of the linear space spanned by $\boldsymbol{\alpha}^{(1)},\ldots,\boldsymbol{\alpha}^{(2K)}$ is

at most $K-1$. So we can find $K-1$ coefficients such that all other coefficients are their linear combinations. Without loss of generality, assume those coefficients are $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K-1)}$. Define the support of a vector to be the entries where it is nonzero. For $\boldsymbol{\alpha}^{(K)}, \ldots, \boldsymbol{\alpha}^{(2K)}$, there will be one coefficient whose support is contained in the union of all the other coefficients. If this is not the case, each coefficient can be mapped to one entry which is only contained in its own support but not any support of other coefficients. But there are $K+1$ coefficient and only K entries, which leads to a contradiction. Without loss of generality, assume that coefficient is $\boldsymbol{\alpha}^{(K)}$. Now we will show that $(\underbrace{1, \ldots, 1}_{K}, 0, \ldots, 0) \notin \mathcal{F}^{(2K)}(\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(2K)})$. Since $f_c(\boldsymbol{\alpha}^{(i)}) = 1$ for $i = 1, \ldots, K-1$, we have

$$\sum_j c_j \boldsymbol{\alpha}_j^{(i)} = 0 \quad \forall\, i = 1, \ldots, K-1.$$

Because $\boldsymbol{\alpha}^{(K)}, \ldots, \boldsymbol{\alpha}^{(2K)}$ are linear combinations of $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K-1)}$, we know

$$\sum_j c_j \boldsymbol{\alpha}_j^{(i)} = 0 \quad \forall\, i = K, \ldots, 2K.$$

If $f_c(\boldsymbol{\alpha}^{(i)}) = 0$ for $i = K+1, \ldots, 2K$, it means

$$\sum_j \left(c_j \boldsymbol{\alpha}_j^{(i)}\right)^2 = 0 \quad \forall\, i = K+1, \ldots, 2K,$$

which means the support of c does not overlap with the support of $\boldsymbol{\alpha}^{(K+1)}, \ldots, \boldsymbol{\alpha}^{(2K)}$. However, the support of $\boldsymbol{\alpha}^{(K)}$ is contained in the union of the supports of $\boldsymbol{\alpha}^{(K+1)}, \ldots, \boldsymbol{\alpha}^{(2K)}$. That means $f_c(\boldsymbol{\alpha}^{(K)} = 0$ not 1, a contradiction.

Then by the classic statistical learning theory, for example, see Theorem 4.1 in [93], we know (A.37) holds true.

Now comes the second major step: we want to show that

$$A(\epsilon, \rho_1, \rho_2) \Rightarrow \sup_{c_1, \ldots, c_K} \frac{1}{n} \sum_{i=1}^n f_c(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{2L\rho_2}.$$

Then, using (A.37), we get the desired conclusion.

For any $\epsilon, \rho_1, \rho_2 > 0$, if $\boldsymbol{D} \neq \boldsymbol{D}^*$ is a local min with sharpness at least ϵ and eigenvalue$(\boldsymbol{D}) \in [\rho_1, \rho_2]$, then $\sup_{c_1, \ldots, c_K} \frac{1}{n} \sum_{i=1}^n f_c(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{L\rho_2}$. Since $\boldsymbol{D} \neq \boldsymbol{D}^*$ up to sign-permutation ambiguity, at least one row of $\boldsymbol{D}^{-1}\boldsymbol{D}^*$ contains two nonzero elements. Without loss of generality, assume the k-th row of $\boldsymbol{D}^{-1}\boldsymbol{D}^*$, denoted as $c^{(k)}$, has at least two nonzero entries. We will prove that it satisfies the desired condition:

$$\frac{1}{n} \sum_{i=1}^n f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{2L\rho_2}.$$

Recall that $\boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$ for $i = 1, \ldots, n$. Because $|||\boldsymbol{D}^{-1}|||_2 \leq \rho_1^{-1}$ and $\|\boldsymbol{x}^{(i)}\|_2$ is bounded by L, for any vector w such that $\|w\|_2 = 1$, we have $|\sum_j w_j \boldsymbol{\beta}_j^{(i)}| \leq L\rho_1^{-1}$ by Cauchy inequality. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \geq & \frac{\rho_1}{L}\max_i\{\Big|\sum_j w_j\boldsymbol{\beta}_j^{(i)}\Big|\}\frac{1}{n}\sum_i f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \\ \geq & \frac{\rho_1}{L}\frac{1}{n}\sum_i |\sum_j w_j\boldsymbol{\beta}_j^{(i)}| f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \\ = & \frac{\rho_1}{L}\frac{1}{n}\sum_i |\sum_j w_j\boldsymbol{\beta}_j^{(i)}| \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j (c_j^{(k)}\boldsymbol{\alpha}_j^{(i)})^2 > 0). \end{aligned} \tag{A.38}$$

Note that this inequality holds for any w with unit norm. Recall that $c_j^{(k)}$ has at least two non-zero entries. Thus, for all the i's such that $\sum_j (c_j^{(k)}\boldsymbol{\alpha}_j^{(i)})^2 = 0$, $\boldsymbol{\alpha}_j^{(i)}$ must satisfy at least two linear constraints, which implies the corresponding $\boldsymbol{\beta}^{(i)}$'s must lie in a linear subspace of dimensions at most $K-2$. Therefore, we can always select w such that $w_k = 0$ and for any i such that $\sum_j (c_j^{(k)}\boldsymbol{\alpha}_j^{(i)})^2 = 0$, we have $\sum_j w_j\boldsymbol{\beta}_j^{(i)} = 0$. Then, by using this specific w (this w satisfies $\|w\|_2 = 1, w_k = 0$, and for any $i = 1, \ldots, n$, $\sum_j w_j\boldsymbol{\beta}_j^{(i)}\mathbf{1}\left(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j \left(c_j^{(k)}\boldsymbol{\alpha}_j^{(i)}\right)^2 = 0\right) = 0$), we have:

$$\frac{\rho_1}{L}\frac{1}{n}\sum_i |\sum_j w_j\boldsymbol{\beta}_j^{(i)}| \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j (c_j^{(k)}\boldsymbol{\alpha}_j^{(i)})^2 > 0) = \frac{\rho_1}{L}\frac{1}{n}\sum_i |\sum_j w_j\boldsymbol{\beta}_j^{(i)}| \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0). \tag{A.39}$$

Using the parametrization in Proposition 5.2.1, for $t > 0$ sufficiently small, $t \cdot w$ can map to a $\boldsymbol{D}'$ in the neighborhood of $\boldsymbol{D}$ such that $|||\boldsymbol{D}' - \boldsymbol{D}|||_F \geq \rho_1^2 |||\boldsymbol{D}'^{-1} - \boldsymbol{D}^{-1}|||_F \geq \rho_1^2 \cdot t \cdot \|w^T\boldsymbol{D}^{-1}\|_2 \geq \frac{\rho_1^2 t}{\rho_2}$. Because $\boldsymbol{D}$ is sharp local minimum with sharpness at least ϵ, we have

$$\frac{1}{n}\sum_{i=1}^{n} \|(\boldsymbol{D}')^{-1}\boldsymbol{x}^{(i)}\|_1 - \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}\|_1 \geq \epsilon|||\boldsymbol{D}' - \boldsymbol{D}|||_F + o(|||\boldsymbol{D}' - \boldsymbol{D}|||_F).$$

By Lemma A.4.1, we know the left hand side of the above inequality is equivalent to

$$|||(\boldsymbol{D}')^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} - \mathrm{tr}(((\boldsymbol{D}')^{-1}\boldsymbol{D})^T B(\boldsymbol{\beta}, \boldsymbol{D}^T\boldsymbol{D}) \geq \epsilon|||\boldsymbol{D}' - \boldsymbol{D}|||_F + o(|||\boldsymbol{D}' - \boldsymbol{D}|||_F).$$

Without loss of generality, we could select $\boldsymbol{D}'$ (or $-\boldsymbol{D}'$) such that

$$\mathrm{tr}(((\boldsymbol{D}')^{-1}\boldsymbol{D})^T B(\boldsymbol{\beta}, \boldsymbol{D}^T\boldsymbol{D}) \geq 0.$$

This means the above inequality can be further rewritten as

$$|||(\boldsymbol{D}')^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} \geq \frac{\rho_1^2\epsilon}{\rho_2}t + o(t).$$

Note that

$$\begin{aligned}|||(\boldsymbol{D}')^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{k'}\left|\boldsymbol{\beta}_{k'}^{\prime(i)}\right|\cdot\mathbf{1}(\boldsymbol{\beta}_{k'}^{(i)}=0)\\ =& t\cdot\frac{1}{n}\sum_{i}|\sum_{j}w_j\boldsymbol{\beta}_j^{(i)}|\mathbf{1}(\boldsymbol{\beta}_k^{(i)}=0).\end{aligned}$$

That means that when t is small,

$$\frac{1}{n}\sum_{i}|\sum_{j}w_j\boldsymbol{\beta}_j^{(i)}|\mathbf{1}(\boldsymbol{\beta}_k^{(i)}=0) > \frac{\rho_1^2\epsilon}{2\rho_2}. \tag{A.40}$$

Combining (A.38), (A.39), and (A.40), we complete the proof. □

Proof of Proposition 5.1.1. 1) $\leftrightarrow$ 2): First observe that 2) is equivalent to the property that the directional derivative of the optimization (5.1) at $\mathbb{I}_k$ along any direction is always positive. By Theorem 4.1.1, we know 1) is equivalent to

$$|||B(\boldsymbol{\beta}, M)|||_{\boldsymbol{\alpha}}^{*} < 1.$$

Because of the definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, this condition is equivalent to for any $k = 1, \ldots, K$, and $\delta_{k,j} \in \mathbb{R}$ for $j = 1, \ldots, K$, $j \neq k$ such that $\sum_{j\neq k}\delta_{k,j}^2 > 0$,

$$\left|\sum_j \delta_{k,j}\beta_j\right|\mathbf{1}(\beta_k = 0) + \sum_{k,j}\delta_{k,j}\beta_j\text{sign}(\beta_k) - M_{k,j}\mathbb{E}|\beta_k| > 0.$$

On the other hand, the left hand side is exactly the directional derivative of the optimization (5.1) at $\mathbb{I}_j$ along direction $(\delta_1, \ldots, \delta_K)$. Because every directional derivative is strictly positive, $\mathbb{I}_k$ is a sharp local minimum of the optimization.

2) $\leftrightarrow$ 3). We have already shown that 2) is equivalent to

$$\left|\sum_j \delta_{k,j}\beta_j\right|\mathbf{1}(\beta_k = 0) + \sum_{k,j}\delta_{k,j}\beta_j\text{sign}(\beta_k) + M_{k,j}\mathbb{E}|\beta_k| > 0.$$

3) is equivalent to

$$\left|\sum_j \delta_{k,j}\beta_j\right|\mathbf{1}(\beta_k = 0) + \sum_{k,j}\delta_{k,j}\beta_j\text{sign}(\beta_k) + \tilde{M}_{k,j}\mathbb{E}|\beta_k| \geq 0.$$

for any $|\tilde{M}_{k,h} - M_{k,h}| \leq \rho$. These two are clearly equivalent for a sufficiently small ρ. □

Proof of Proposition 5.2.1. Without loss of generality, we only need to show $\|Q_j^{-1}\|_2 = 1$ for any $j = 1, \ldots, K$ when $k = 1$. We can write $Q = \Gamma\boldsymbol{D}^{-1}$, where the matrix Γ is equal to:

$$\Gamma_{h,j} = \begin{cases} w_j & \text{if } h = 1 \\ \sqrt{(w_h - M_{1,h})^2 + 1 - m_h^2} & \text{if } j = h \neq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{A.41}$$

Note that Γ is upper triangle so we can obtain its inverse easily. Then $Q^{-1} = \boldsymbol{D}\Gamma^{-1}$ and

$$\Gamma^{-1}_{h,j} = \begin{cases} 1 & h = 1, j = 1 \\ -w_j/(\sqrt{(w_h - M_{1,h})^2 + 1 - M^2_{1,h}}) & h = 1, j > 1 \\ 1/(\sqrt{(w_h - M_{1,h})^2 + 1 - M^2_{1,h}}) & h > 1, j = h \\ 0 & h > 1, j \neq h \end{cases}$$

Q^{-1}'s first column has the form: $\|Q_1^{-1}\|_2^2 = \|\boldsymbol{D}_1\|_2^2 = 1$. For any other column Q_j^{-1} where $j > 1$, $\|Q_j^{-1}\|_2^2 = \|w_j\boldsymbol{D}_1 - \boldsymbol{D}_j\|_2^2/((w_j - M_{1,j})^2 + 1 - M^2_{1,j}) = 1$. □

Proof of Proposition 5.2.2. Recall

$$f(\boldsymbol{D}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \min(|\boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)}|, \tau).$$

Denote $\boldsymbol{\beta}^{(i)} = (\boldsymbol{D}^{(t,j)})^{-1}\boldsymbol{x}^{(i)}$ and define a new function

$$\tilde{f}(\boldsymbol{D}) = \sum_{j=1}^{K} \sum_{i=1, |\boldsymbol{\beta}_j^{(i)}| \leq \tau}^{n} |\boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)}| + \sum_{i=1, |\boldsymbol{\beta}_j^{(i)}| > \tau}^{n} \tau.$$

Note that for any $\boldsymbol{D}$, $\tilde{f}(\boldsymbol{D})$ is always no smaller than $f(\boldsymbol{D})$, that is, $\tilde{f}(\boldsymbol{D}) \geq f(\boldsymbol{D})$. Also, $\tilde{f}(\boldsymbol{D}^{(t,j)}) = f(\boldsymbol{D}^{(t,j)})$. Because of Proposition 5.2.1, we know the iterate $\boldsymbol{D}^{(t,j+1)}$ in Algorithm 2 is the optimal solution of the following optimization:

$$\begin{aligned} &\text{argmin}_Q \ \tilde{f}(Q^{-1}) \\ &\text{subject to } Q \text{ is parameterized as in Proposition 5.2.1.} \end{aligned}$$

That means $\tilde{f}(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j)})$. Combining the fact that $f(\boldsymbol{D}^{(t,j)}) = \tilde{f}(\boldsymbol{D}^{(t,j)})$ and $f(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j+1)})$, we have $f(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j)}) = f(\boldsymbol{D}^{(t,j)})$. □

Appendix B

Proof of Part II

B.1 Proof of Theorem 8.1.1

Proof. To state the proof of the theorem, we need to define more notations. For a generic set $A \subset [0,1]^p$, with slight abuse of notations, let $N_n(A) = \sum_i \mathbf{1}(\mathbf{x}_i \in A)$ be the number of samples with input features in A, and

$$\mu_n(A) = \frac{\sum_{\mathbf{x}_i \in A} y_i}{N_n(A)}$$

be the average response of those samples. For any feature X_k and $z \in (0,1)$, let $\Delta_{\mathcal{I}}(A,(k,z))$ be the impurity decrease when splitting A into $A \cap \{X_k \le z\}$ and $A \cap \{z < X_k\}$, and $\Delta_{\mathcal{I}}(A,k) = \sup_{0 \le z \le 1} \Delta_{\mathcal{I}}(A,(k,z))$.

The proof of the theorem proceeds in three parts. First, we prove a lemma which gives a tail bound for $\Delta_{\mathcal{I}}(A,k)$. Second, we use the lemma and union bound to derive the upper bound for the expectation of $G_0(T)$. Finally, we use a separate argument based on Gaussian comparison inequalities to obtain the lower bound.

Lemma B.1.1. *For any axis-aligned hyper-rectangle $A \subset [0,1]^p$, $k \notin S$ and $\delta > 0$, we have*

$$\mathbb{P}_{X,\epsilon}(\Delta_{\mathcal{I}}(A,k) \ge \delta \big| N_n(A)) \le 4N_n(A) e^{-\frac{\delta N_n(A)}{4(M+1)^2}}.$$

Proof of Lemma B.1.1. We suppose without loss of generality that $\mathbf{x}_1, \ldots, \mathbf{x}_{N_n(A)} \in A$. For any $z \in [0,1]$, we let

$$A^{\text{left}} = A \cap \{0 \le X_k \le z\}, \quad A^{\text{right}} = A \cap \{z < X_k \le 1\},$$

and introduce the shorthands

$$p^{\text{left}} = \frac{N_n(A^{\text{left}})}{N_n(A)}, \quad p^{\text{right}} = \frac{N_n(A^{\text{right}})}{N_n(A)}, \quad \mu^{\text{left}} = \mu_n(A^{\text{left}}), \quad \mu^{\text{right}} = \mu_n(A^{\text{right}}).$$

Then

$$\begin{aligned}
\Delta_{\mathcal{I}}(A,(k,z)) =& \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A))^2 - \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A^{\text{left}}))^2\,\mathbb{1}(x_{ik}\le z)\\
&- \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A^{\text{right}}))^2\,\mathbb{1}(x_{ik}> z)\\
=& \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A} y_i^2 - \mu_n(A)^2 - p^{\text{left}}\Big(\frac{1}{N_n(A)p^{\text{left}}}\sum_{\mathbf{x}_i\in A} y_i^2\,\mathbb{1}(x_{ik}\le z) - (\mu^{\text{left}})^2\Big)\\
&- p^{\text{right}}\Big(\frac{1}{N_n(A)p^{\text{right}}}\sum_{\mathbf{x}_i\in A} y_i^2\,\mathbb{1}(x_{ik}> z) - (\mu^{\text{right}})^2\Big)\\
=& p^{\text{left}}(\mu^{\text{left}})^2 + p^{\text{right}}(\mu^{\text{right}})^2 - \mu_n(A)^2\\
=& (p^{\text{left}}(\mu^{\text{left}})^2 + p^{\text{right}}(\mu^{\text{right}})^2)(p^{\text{left}}+p^{\text{right}}) - (p^{\text{left}}\mu^{\text{left}} + p^{\text{right}}\mu^{\text{right}})^2\\
=& p^{\text{left}}p^{\text{right}}(\mu^{\text{left}}-\mu^{\text{right}})^2\\
\le& 2p^{\text{left}}p^{\text{right}}[(\mu^{\text{left}}-\mu)^2 + (\mu^{\text{right}}-\mu)^2]\\
\le& 2p^{\text{left}}(\mu^{\text{left}}-\mu)^2 + 2p^{\text{right}}(\mu^{\text{right}}-\mu)^2,
\end{aligned}$$

where

$$\mu = \mathbb{E}[Y|X\in A] = \mathbb{E}[\phi(X)|X\in A].$$

Now suppose without loss of generality that $x_{1k} < x_{2k} < \cdots < x_{nk}$ (otherwise we can reorder the samples by X_k). Since $k\notin S$, X_k is independent of X_S and therefore independent of Y. Thus the distribution of $(y_1,\ldots,y_n)$ does not change after the reordering, i.e.,

$$y_i \overset{i.i.d}{\sim} (\phi(X)|X\in A) + \epsilon.$$

Note that

$$\sup_z\; p^{\text{left}}(\mu^{\text{left}}-\mu)^2 \le \sup_{1\le m\le N_n(A)} \frac{m}{N_n(A)}\left(\frac{1}{m}\sum_{i=1}^m y_i - \mu\right)^2.$$

Note that Y is sub-Gaussian with parameter $M+1$. Therefore, for each $1\le m\le N_n(A)$, by Hoeffding bound,

$$\mathbb{P}\left(\frac{m}{N_n(A)}\left(\frac{1}{m}\sum_{i=1}^m y_i-\mu\right)^2 \ge \delta \,\middle|\, N_n(A)\right) \le 2e^{-(M+1)^2\delta N_n(A)^2/m} \le 2e^{-\frac{\delta N_n(A)}{(M+1)^2}}.$$

Therefore

$$\mathbb{P}\left(\sup_z p^{\text{left}}(\mu^{\text{left}}-\mu)^2 \ge \delta \,\middle|\, N_n(A)\right) \le 2N_n(A)e^{-\frac{\delta N_n(A)}{(M+1)^2}}.$$

By symmetry, the same bound holds for $p^{\text{right}}(\mu^{\text{right}} - \mu)^2$. Therefore

$$\begin{aligned}
&\mathbb{P}(\Delta_{\mathcal{I}}(A,k) \geq \delta \big| N_n(A)) \\
\leq &\mathbb{P}\left(\sup_z p^{\text{left}}(\mu^{\text{left}} - \mu)^2 \geq \delta/2 \big| N_n(A)\right) + \mathbb{P}\left(\sup_z p^{\text{right}}(\mu^{\text{right}} - \mu)^2 \geq \delta/2 \big| N_n(A)\right) \\
\leq &4N_n(A)e^{-\frac{\delta N_n(A)}{4(M+1)^2}},
\end{aligned}$$

and the lemma is proved. □

Proof of the upper bound in Theorem 8.1.1

Without loss of generality, assume that when we split on feature k, the cut is always performed along the direction of k at some data point (and that data point falls into the right sub-tree). Suppose that ϵ_i has unit variance for all i. Let $C = 2\max\{256, 16(M+1)^2\}$. We also assume, without loss of generality, that $m_n \geq 8d_n$. Otherwise, since $G_0(T)$ is, by definition, upper bounded by the sample variance of y, we have

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq \text{Var}(Y) \leq M^2 + 1 \leq 16(M+1)^2 \frac{d_n \log np}{m_n}.$$

To simplify notation, we define $\mathbf{x}_{n+1} = (0, \ldots, 0)$ and $\mathbf{x}_{n+2} = (1, \ldots, 1)$. For any $V \subset [p], \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}$, let

$$A(V, \mathcal{L}, \mathcal{R}) = \{X = (X_1, \ldots, X_p) : x_{\mathcal{L}_i, V_i} \leq X_{V_i} < x_{\mathcal{R}_i, V_i}, 1 \leq i \leq |V|, 0 \leq X_k \leq 1, k \notin V\}$$

be the random axis-aligned hyper-rectangle obtained by splitting on features in V, where the left and right endpoints of the ith feature V_i are determined by $x_{\mathcal{L}_i, V_i}$ and $x_{\mathcal{R}_i, V_i}$. Note that in this definition, we treat $\mathbf{x}_i$ as random variables rather than fixed, and $A(V, \mathcal{L}, \mathcal{R})$ can be the empty set with non-zero probability. Let

$$A(V) = \{A(V, \mathcal{L}, \mathcal{R}) | \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}\}$$

be all axis-aligned hyper-rectangles obtained by splitting on features in V. For any $d \leq d_n$, let

$$A_d = \cup_{|V|=d} A(V)$$

be the collection of all possible subsets of $[0,1]^p$ obtained by splitting on d features.

Fix $\delta > \frac{96M^2 d_n}{m_n}$. We will first show that

$$\begin{aligned}
&\mathbb{P}_{X,\epsilon}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n \delta}{N_n(A)} \text{ and } N_n(A) \geq m_n\right) \\
\leq &5(np)^{d+1} \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right).
\end{aligned} \tag{B.1}$$

Note that for any two events C_1 and C_2, the inequality $\mathbb{P}(C_1 \cap C_2) \leq \mathbb{P}(C_1|C_2)$ always holds. Therefore, for any hyper-rectangle A, we have

$$\begin{aligned} &\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)} \text{ and } N_n(A) \geq m_n\right) \\ \leq &\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)} \middle| N_n(A) \geq m_n\right) \end{aligned} \tag{B.2}$$

To simplify notation, we will drop the conditional event $N_n(A) \geq m_n$ in the remainder of the proof of the upper bound, unless stated otherwise.

Fix $V \subset [p], \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}$, and $k \notin S$. Conditional on samples in $\mathcal{L}$ and $\mathcal{R}$, we would like to apply Lemma B.1.1 to $A(V,\mathcal{L},\mathcal{R})$ and k. The only problem is that there are now samples on the boundary of $A(V,\mathcal{L},\mathcal{R})$, namely those in $\mathcal{L}$ and $\mathcal{R}$. Let $\mathbf{x}_{\mathcal{L}} = \{\mathbf{x}_i\}_{i\in\mathcal{L}}$ and $\mathbf{x}_{\mathcal{R}} = \{\mathbf{x}_i\}_{i\in\mathcal{R}}$. Conditional on $\mathbf{x}_{\mathcal{L}}$, $\mathbf{x}_{\mathcal{R}}$ and $N_n(A(V,\mathcal{L},\mathcal{R}))$, and on the random variable $X \in A(V,\mathcal{L},\mathcal{R})$, X is uniformly distributed in $A(V,\mathcal{L},\mathcal{R})$. For a set A, we let A° be the interior of A and let $\bar{A}$ be the boundary of A. Since $m_n \geq 8d_n$,

$$\frac{N_n(A^\circ(V,\mathcal{L},\mathcal{R}))}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n - 2d_n}{m_n} \geq \frac{3}{4}.$$

By Lemma B.1.1, we have

$$\begin{aligned} &\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\middle| \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{R}}, N_n(A(V,\mathcal{L},\mathcal{R}))\right) \\ \leq &4N_n(A^\circ(V,\mathcal{L},\mathcal{R}))\exp\left(-\frac{\delta m_n N_n(A^\circ(V,\mathcal{L},\mathcal{R}))}{12(M+1)^2 N_n(A(V,\mathcal{L},\mathcal{R}))}\right) \\ \leq &4n\exp\left(-\frac{\delta m_n}{16(M+1)^2}\right) \end{aligned} \tag{B.3}$$

for large n. Since the right hand side does not depend on $\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{R}}, N_n(A(V,\mathcal{L},\mathcal{R}))$, we can take expectation with respect to them, and obtain

$$\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right) \leq 4n\exp\left(-\frac{\delta m_n}{16(M+1)^2}\right) \tag{B.4}$$

On the other hand, we have the inequality

$$\begin{aligned} \Delta_{\mathcal{I}}(A(V,\mathcal{L},\mathcal{R}),k) &\leq \Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) + \frac{\sum_{i\in\mathcal{L},\mathcal{R}}(y_i - \mu_n(A(V,\mathcal{L},\mathcal{R})))^2}{N_n(A(V,\mathcal{L},\mathcal{R}))} \\ &\leq \Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) + \frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2(y_i^2 + \mu_n(A(V,\mathcal{L},\mathcal{R}))^2)}{N_n(A(V,\mathcal{L},\mathcal{R}))}. \end{aligned} \tag{B.5}$$

We have

$$\begin{aligned}
&\mathbb{P}_{X,\epsilon}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2y_i^2}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4(f^2(\mathbf{x}_i)+\epsilon_i^2)}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4(f^2(\mathbf{x}_i)+\epsilon_i^2)}{m_n} \geq \frac{\delta}{3}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4M^2+4\epsilon_i^2}{m_n} \geq \frac{\delta}{3}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i=1}^{2d_n}(\epsilon_i^2-1)}{m_n} \geq \frac{\delta}{16N_n(A^\circ(V,\mathcal{L},\mathcal{R})}\right) \leq \exp(-\frac{\delta m_n}{256}),
\end{aligned} \tag{B.6}$$

for large n, where the fourth inequality holds because $\delta \geq 96M^2d_n/m_n$, and the last inequality follows from the well-known tail bound

$$\mathbb{P}\left(\left|\frac{1}{d}\chi_d^2 - 1\right| \geq \delta_0\right) \leq 2e^{-d\delta_0^2/8}$$

for χ_d^2 random variable and $\delta_0 < 1$. To upper bound $\mu_n(A(V,\mathcal{L},\mathcal{R}))$, note that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2\mu_n(A(V,\mathcal{L},\mathcal{R}))^2}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(|\mu_n(A(V,\mathcal{L},\mathcal{R}))| \geq \sqrt{\frac{\delta m_n}{6d_n}}\right)\\
\leq&\mathbb{P}\left(\left|\frac{1}{N_n(A(V,\mathcal{L},\mathcal{R}))}\sum_{i=1}^{N_n(A(V,\mathcal{L},\mathcal{R}))}\epsilon_i\right| \geq \sqrt{\frac{\delta m_n}{6d_n}} - M\right)\\
\leq&2\exp\left(-\frac{1}{2}m_n(\sqrt{\frac{\delta m_n}{6d_n}} - M)^2\right)\\
\leq&2\exp\left(-\frac{\delta m_n}{4}\right),
\end{aligned} \tag{B.7}$$

where the last inequality follows from $m_n \geq 8d_n$ and $\delta \geq 96M^2d_n/m_n$. Combining Equations (B.4), (B.5), (B.7), we have

$$\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right) \leq 5n\exp\left(-\frac{\delta m_n}{\max\{16(M+1)^2,256\}}\right) \tag{B.8}$$

for any $V \subset [p], |V| = d, \mathcal{L},\mathcal{R} \in [n+2]^{|V|}$, and $k \notin S$. Note that the set A_d has cardinality

$$|A_d| = \binom{p}{d}(2(n+2))^d \leq \left(\frac{pn}{d}\right)^d$$

for large n. Therefore by union bound,

$$\mathbb{P}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}\right) \leq 5np|A_d| \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right)$$
$$\leq 5(np)^{d+1} \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right). \tag{B.9}$$

Suppose that $\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}$ for all $A \in \cup_{d\leq d_n} A_d$ and $k \notin S$, then for any $T \in T_n(m_n, d_n)$,

$$G_0(T) \leq \sum_{t:v(t)\notin S} \frac{N_n(t)}{n}\frac{m_n\delta}{N_n(t)} \leq \delta\frac{m_n|I(t)|}{n} \leq \delta,$$

where the last inequality follows since $|I(t)|+1$ is the total number of leaf nodes in T, and each leaf node contains at least m_n samples. Therefore

$$\mathbb{P}_{X,\epsilon}\left(\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T) \geq \delta\right) \leq \sum_{d=1}^{d_n} \mathbb{P}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}\right)$$
$$\leq \sum_{d=1}^{d_n} 5(np)^{d+1} \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right) \tag{B.10}$$
$$\leq 10(np)^{d_n+1} \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right)$$

for any $\delta > \frac{96M^2d_n}{m_n}$. Recall that $C = 2\max\{256, 16(M+1)^2\}$. Note that $\frac{Cd_n\log(np)}{m_n} \geq \frac{96M^2d_n}{m_n}$ for large n. Integrating over δ, we have

$$\begin{aligned}
&\mathbb{E}_{X,\epsilon}\left[\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T)\right] \\
&\leq \frac{3d_n\log(np)}{2m_n} + \mathbb{E}_{X,\epsilon}\left[\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T)\mathbf{1}(\delta \geq \frac{3d_n\log(np)}{2m_n})\right] \\
&\leq \frac{3d_n\log(np)}{2m_n} + \int_{\frac{3d_n\log(np)}{2m_n}}^{\infty} \mathbb{P}_{X,\epsilon}\left(\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T) \geq \delta\right) d\delta \\
&\leq \frac{Cd_n\log(np)}{m_n}.
\end{aligned} \tag{B.11}$$

This completes the proof of the upper bound.

Proof of the lower bound in Theorem 8.1.1

For the lower bound, let

$$d_n = \max\{d : 2^{d+1}m_n < n\}, \tag{B.12}$$

and consider a balanced, binary decision tree T constructed in the following way:

1. At each node on the first $d_n - 1$ levels of the tree, we split on feature X_1, at the mid-point of X_1's side of the rectangle corresponding to the node.

2. At each node on the d_nth level, we look at the remaining $p-1$ features, and split on the feature that maximizes the decrease in impurity.

In the following proof, we will lower bound $G_0(T)$ by the sum of impurity reduction on the d_nth level alone. For $t = 1, \ldots, 2^{d_n-1}$, let

$$R_t = \left\{ \frac{t-1}{2^{d_n-1}} \leq X_1 < \frac{t}{2^{d_n-1}} \right\}.$$

be the hyper-rectangle corresponding to the tth node on the d_nth level. Applying Chernoff's inequality, we have

$$\mathbb{P}\left(\left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \geq \frac{1}{3 \cdot 2^{d_n-1}} \right) \leq 2 \exp\left(- \frac{n}{27 \cdot 2^{d_n-1}} \right).$$

Let

$$B_1 = \left\{ \left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \leq \frac{1}{3 \cdot 2^{d_n-1}} \text{ for all } t \right\}$$

be the event that each node on the d_nth level contains at least $\frac{2}{3}\frac{n}{2^{d_n-1}}$, but no more than $\frac{4}{3}\frac{n}{2^{d_n-1}}$ samples. Then

$$\mathbb{P}(B_1^c) \leq \sum_{t=1}^{2^{d_n-1}} \mathbb{P}\left(\left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \geq \frac{1}{3 \cdot 2^{d_n-1}} \right) \leq 2^{d_n} \exp\left(- \frac{n}{27 \cdot 2^{d_n-1}} \right), \tag{B.13}$$

and conditional on B_1,

$$\frac{8}{3} m_n \leq \frac{2}{3}\frac{n}{2^{d_n-1}} \leq N_n(R_t) \leq \frac{4}{3}\frac{n}{2^{d_n-1}} \leq \frac{32}{3} m_n. \tag{B.14}$$

We define

$$R_t^l(k) = R_t \cap \left\{ 0 \leq X_k < \frac{1}{2} \right\}$$

and

$$R_t^r(k) = R_t \cap \left\{ \frac{1}{2} \leq X_k < 1 \right\}$$

and use R_t^l, R_t^r as shorthand when k is fixed. For each $t = 0, 1, \ldots, 2^d - 1$, by Equation

$$\Delta_{\mathcal{I}}(R_t, k) \geq \Delta_{\mathcal{I}}(R_t, (k, 1/2)) = \frac{N_n(R_t^l)}{N_n(R_t)} \frac{N_n(R_t^r)}{N_n(R_t)} (\mu_n(R_t^l) - \mu_n(R_t^r))^2$$

Let

$$\eta_k = \mu_n(R_t^l) - \mu_n(R_t^r)$$

Conditional on $N_n(R_t^l)$ and $N_n(R_t^r)$, $\eta = (\eta_2, \ldots, \eta_p)$ are jointly Gaussian with zero mean. To lower bound the impurity decrease at the tth node on the d_nth level, we use a Gaussian comparison argument to obtain a lower bound for $\sup_k |\eta_k|$, which requires us to calculate the covariance matrix of η. For any $2 \le k_1, k_2 \le p$, let us further define

$$R_t^{ll}(k_1,k_2) = R_t \cap \left\{0 \le X_{k_1} < \frac{1}{2}\right\} \cap \left\{0 \le X_{k_2} < \frac{1}{2}\right\};$$

$$R_t^{lr}(k_1,k_2) = R_t \cap \left\{0 \le X_{k_1} < \frac{1}{2}\right\} \cap \left\{\frac{1}{2} \le X_{k_2} < 1\right\};$$

$$R_t^{rl}(k_1,k_2) = R_t \cap \left\{\frac{1}{2} \le X_{k_1} < 1\right\} \cap \left\{0 \le X_{k_2} < \frac{1}{2}\right\};$$

$$R_t^{rr}(k_1,k_2) = R_t \cap \left\{\frac{1}{2} \le X_{k_1} < 1\right\} \cap \left\{\frac{1}{2} \le X_{k_2} < 1\right\}.$$

As before, we write $R_t^{ll}, R_t^{lr}, R_t^{rl}$ and R_t^{rr} as shorthand when k_1, k_2 are fixed. Conditional on $N_n(R_t)$, the samples falling into the hyper-rectangle R_t are uniformly distributed in R_t. Therefore we know from Chernoff's inequality that

$$\mathbb{P}\left(\left|\frac{N_n(R_t^{ll})}{N_n(R_t)} - \frac{1}{4}\right| \ge \frac{1}{16}\right) \le 2\exp\left(-\frac{N_n(R_t)}{48}\right)$$

for any k_1 and k_2, and that the same results hold for R_t^{lr}, R_t^{rl} and R_t^{ll} as well. Let

$$B_2 = \left\{\max_{\omega \in \{ll,lr,rl,rr\}} \left|\frac{N_n(R_t^{\omega}(k_1,k_2))}{N_n(R_t)} - \frac{1}{4}\right| \le \frac{1}{16}, \text{ for all } 1 \le t \le 2^{d_n-1}, 2 \le k_1 < k_2 \le p\right\}.$$

Then

$$\mathbb{P}(B_2^c) \le 2^{d_n} p^2 \exp\left(-\frac{N_n(R_t)}{48}\right), \tag{B.15}$$

and

$$\mathbb{P}(B_1 \cap B_2) \ge 1 - 2^{d_n+1} p^2 \exp\left(-\frac{N_n(R_t)}{48}\right) \ge 1 - 2^{d_n+1} p^2 \exp\left(-\frac{m_n}{18}\right) \ge \frac{8}{9} \tag{B.16}$$

for n large enough (under the condition that $m_n \ge 36\log p + 18\log n$). Conditional on the event B_2,

$$N_n(R_t^l) \ge N_n(R_t^{ll}) + N_n(R_t^{lr}) \ge \frac{3}{16}N_n(R_t) + \frac{3}{16}N_n(R_t) \ge \frac{3}{8}N_n(R_t),$$

for any $1 \le t \le 2^{d_n-1}$ and $2 \le k \le p$, and the same holds for $N_n(R_t^r)$. Therefore,

$$\mathrm{Var}(\eta_k) = \frac{1}{N_n(R_t^l)} + \frac{1}{N_n(R_t^r)} \ge \frac{3}{4N_n(R_t)} \tag{B.17}$$

$$\mathrm{Cov}(\eta_{k_1}, \eta_{k_2}) = \frac{1}{N_n(R_t^{ll})} + \frac{1}{N_n(R_t^{rr})} - \frac{1}{N_n(R_t^{lr})} - \frac{1}{N_n(R_t^{rl})} \le \frac{1}{4N_n(R_t)}. \tag{B.18}$$

Consider $\tilde{\eta}_2, \ldots, \tilde{\eta}_p$ with

$$\mathbb{E}\tilde{\eta}_k = 0, \mathrm{Var}(\tilde{\eta}_k) = \frac{3}{4N_n(R_t)}$$

and

$$\mathrm{Cov}(\tilde{\eta}_{k_1}, \tilde{\eta}_{k_2}) = \frac{1}{4N_n(R_t)}.$$

Then conditional on $B_1 \cap B_2$, by Sudakov-Fernique lemma, we have

$$\mathbb{E}_\epsilon[\max_k \eta_k | B_1 \cap B_2] \ge \mathbb{E} \max_k \tilde{\eta}_k \ge \sqrt{\frac{\log p}{N_n(R_t)}} \ge \sqrt{\frac{3\log p}{32 m_n}},$$

and the lower bound

$$\min\{N_n(R_t^l), N_n(R_t^r)\} \ge \frac{3}{8} N_n(R_t) \ge m_n,$$

for any k, t. where the last inequality follows from Equation (B.14). Therefore, conditional on $B_1 \cap B_2$ the minimum leaf size is lower bounded by m_n. Finally

$$\begin{aligned}
\mathbb{E}_{X,\epsilon}\left[\sup_{T \in \mathcal{T}_n(m_n)} G_0(T)\right] &\ge \mathbb{E}_{X,\epsilon}\left[\sup_{T \in \mathcal{T}_n(m_n)} G_0(T)\mathbf{1}_{B_1 \cap B_2}\right] \\
&\ge \mathbb{E}_X\left[\sum_t \frac{N_n(R_t)}{n} \mathbb{E}_\epsilon\left[\max_k \Delta_{\mathcal{I}}(R_t, k)\mathbf{1}_{B_1 \cap B_2}\right]\right] \\
&\ge \mathbb{E}_X \sum_t \frac{N_n(R_t)}{n} (\frac{3}{8})^2 (\mathbb{E}_\epsilon \max_k \eta_k^2 \mathbf{1}_{B_1 \cap B_2}) \\
&\ge \frac{9}{64} \frac{3\log p}{32 m_n} \mathbb{P}(B_1 \cap B_2) \\
&\ge \frac{1}{80} \frac{\log p}{m_n}
\end{aligned} \tag{B.19}$$

when n is large enough, and the lower bound is proved. This concludes the whole proof.

□

B.2 Proof of Theorem 9.4.1

Proof of the population case – desirable features

Recall from Section 9.2 that there are three different sources of randomness:

1. ($\mathcal{D}$) the randomness of the data $\mathcal{D}$,
2. (T) the randomness of the tree T, given the data $\mathcal{D}$,
3. ($\mathcal{P}$) the randomness of the randomly selected path $\mathcal{P}$, given the tree T.

Note that, although the random path $\mathcal{P}$ depends on all three sources of randomness (the data randomness, the tree randomness, and the additional path randomness), when we condition on the tree T, then the random path $\mathcal{P}$ is independent of the data $\mathcal{D}$. In the first part of the proof, we will only consider the last two sources of randomness, namely, from the random tree and from the randomly selected path on the tree. Also, recall that we define $S_j^-, S_j^+ \subset [p] \times \{-1, +1\}$ as the features in $S_j \subset [p]$ with $-$ and $+$ sign, respectively, that is

$$S_j^- = \{(k, -1) \ : \ k \in S_j\} \subset [p] \times \{-1, +1\}, \tag{B.20}$$
$$S_j^+ = \{(k, +1) \ : \ k \in S_j\} \subset [p] \times \{-1, +1\}. \tag{B.21}$$

For each node t in a tree T, define $\dot{\mathrm{F}}^{\pm}(t)$ to be the set of signed features used by the parents of t in T and $\dot{\mathrm{F}}(t)$ to be the corresponding (unsigned) features. For any feature j, $(j, -)$ and $(j, +)$ can appear together in $\dot{\mathrm{F}}^{\pm}(t)$. Furthermore, let $\mathrm{F}^{\pm}(t)$ be a subset of $\dot{\mathrm{F}}^{\pm}(t)$ by only including the signed feature that corresponds to the first split of the feature if a feature appeared multiple times in the path. As a result, for any feature j, at most one of $(j, +)$ and $(j, -)$ can appear in $\dot{\mathrm{F}}^{\pm}(t)$. Define $\mathrm{F}(t)$ to be the set of (unsigned) features in $\mathrm{F}^{\pm}(t)$. Because $\dot{\mathrm{F}}^{\pm}(t)$ and $\mathrm{F}^{\pm}(t)$ only differ in terms of feature signs, they correspond to the same set of features, i.e., $\dot{\mathrm{F}}(t) = \mathrm{F}(t)$. Conditioned on a tree T, at every node t of T we now define the set of *desirable* features with respect to the LSS model as follows.

Definition B.2.1 (Desirable features)**.** *Define the desirable feature set $U(t) \subset [p]$ to be*

$$U(t) \triangleq \left\{ k \in [p] \,\middle|\, \exists\, j \in [J] \ s.t. \ k \in S_j, \ S_j^+ \cap \mathrm{F}(t) = \emptyset \ and \ (k, -1) \notin \mathrm{F}^{\pm}(t) \right\}. \tag{B.22}$$

Note that the set of desirable features $U(t)$ at a node t is only defined w.r.t. some particular LSS model. In particular, it depends on the basic signed interactions $S_1^-, \ldots, S_J^-$. Hence, for a given tree T with node t, $U(t)$ is an oracle set, which cannot be computed from data. The way to think about $U(t)$ is that it corresponds exactly to those set of features which would yield some impurity decrease if the tree was grown by seeing the full data distribution $P(X, Y)$ and hence, making every split at the correct split point. Moreover, denote t_{leaf} to be

the leaf node of $\mathcal{P}$ and we define $\mathcal{F}$ to be the desirable signed features of $F(t_{\text{leaf}})$. That is, the signed features k_t where for the node t on the path $\mathcal{P}$ we have $k_t \in U(t)$, i.e.,

$$\mathcal{F}(\mathcal{P}) \triangleq \{(k_t, b_t) \in F(t_{\text{leaf}}) \,\Big|\, k_t \in U(t),\ t_{\text{leaf}} \text{ is leaf node of } \mathcal{P}\ \} \subset [p] \times \{-1, +1\} \tag{B.23}$$

For notation simplicity, we use $\mathcal{F}$ as the shorthand of $\mathcal{F}(\mathcal{P})$.

Further, we define the event Ω_0 to be that the desirable features are exhausted at the leaf node:

$$\Omega_0 \triangleq \{U(t_{\text{leaf}}) = \emptyset \text{ for the leaf node } t_{\text{leaf}} \text{ of } \mathcal{P}\}. \tag{B.24}$$

With these definitions we get the following lemma.

Lemma B.2.1. *For the event Ω_0 in (B.24) it holds true that*

$$\Omega_0 \subset \bigcap_{j \in [J]} \{S_j^- \subset \mathcal{F}\} \cup \{S_j^+ \cap \mathcal{F} \neq \emptyset\}, \tag{B.25}$$

with $\{S_j^- \subset \mathcal{F}\} \cap \{S_j^+ \cap \mathcal{F} \neq \emptyset\} = \emptyset$.

Proof. For an arbitrary interaction $j \in [J]$, it follows from the definition of $U(t)$ that Ω_0 implies either $S_j^+ \cap \mathrm{F}^{\pm}(t_{\text{leaf}}) \neq \emptyset$ or $S_j^- \subset \mathrm{F}^{\pm}(t_{\text{leaf}})$. First, consider $S_j^+ \cap \mathrm{F}^{\pm}(t_{\text{leaf}}) \neq \emptyset$. Let $(k, +1) \in S_j^+ \cap \mathrm{F}^{\pm}(t_{\text{leaf}})$ be the signed feature in $S_j^+ \cap \mathrm{F}^{\pm}(t_{\text{leaf}})$ that appears first on the path. Then, because $\mathrm{F}^{\pm}(t)$ only considers the signed features when they first appear in a path, we have that $(k, +1)$ was desirable and thus, $(k, +1) \in \mathcal{F}$, i.e., $S_j^+ \cap \mathcal{F} \neq \emptyset$. Second, consider $S_j^- \subset \mathrm{F}^{\pm}(t_{\text{leaf}})$. Then for any $(k, -1) \in S_j^-$, by definition of $F(t)$ we have that no S_j^+ feature appeared on the path before $(k, -1)$ and hence, $(k, +1) \in \mathcal{F}$, i.e., $S_j^- \subset \mathcal{F}$. Finally, recall that by definition of $\mathcal{F}$ both conditions in (B.25) can never happen at the same time. □

Recall that $\mathcal{F}$ is defined in (B.23) and $\Omega_0 \in \sigma(D, T, \mathcal{P})$ is defined in (B.24). Define

$$C_{\text{root}}(\mathcal{D}) \triangleq \min_{k \in \cup_{j=1}^{J} S_j} P_T\left(t_{\text{root}} \text{ splits on feature } k \mid \mathcal{D}\right). \tag{B.26}$$

We state the population version of our main results below.

Theorem B.2.1. *For all $\tilde{S}^{\pm} \subset [p] \times \{-1, +1\}$ with $\tilde{s} = |\tilde{S}|$ we have that almost surely*

$$P_{\mathcal{P}}\left(\tilde{S}^{\pm} \subset \mathcal{F} \,\Big|\, T, \mathcal{D}\right) \leq 0.5^{\tilde{s}} \tag{B.27}$$

and if $\tilde{S}^{\pm}$ is a union interaction as in Definition 9.1.1 then almost surely

$$P_{\mathcal{P}}\left(\tilde{S}^{\pm} \subset \mathcal{F} \,\Big|\, T, \mathcal{D}\right) \geq 0.5^{\tilde{s}} - P_{\mathcal{P}}\left(\Omega_0^c \mid T, \mathcal{D}\right). \tag{B.28}$$

Moreover, if $\tilde{S}^{\pm}$ is not a union interaction then

$$P_{(\mathcal{P},T)}\left(\tilde{S}^{\pm} \subset \mathcal{F} \,\Big|\, \mathcal{D}\right) \leq 0.5^{\tilde{s}}(1 - C_{\text{root}}(\mathcal{D})/2). \tag{B.29}$$

Proof of Theorem B.2.1. **Proof of** (B.27): Recall that the path $\mathcal{P}$ corresponding to $\mathcal{F}$ is selected in such a way: one starts at the root node t_{root} and then randomly follows the paths in the tree either to the plus ($B = +1$) or to the minus ($B = -1$) direction with probability 0.5 according to the Bernoulli coin flips in $\mathcal{B}$.

For any feature $k \in [p]$, let B^k be the Bernoulli random variable we draw when k appears for the first time on $\mathcal{P}$. Note that when $(k, -1) \in \mathcal{F}$, $B^k = -1$. Similar, when $(k, +1) \in \mathcal{F}$, this implies that $B^k = +1$. Consequently, for any $\tilde{S}^{\pm} = \{(k_1, b_1), \ldots, (k_{\tilde{s}}, b_{\tilde{s}})\} \subset [p] \times \{-1, +1\}$ we have that

$$\{\tilde{S}^{\pm} \subset \mathcal{F}\} \subset \{B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}}\} \tag{B.30}$$

and hence

$$P\left(\tilde{S}^{\pm} \subset \mathcal{F}\right) \leq P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}}\right) = 0.5^{\tilde{s}}, \tag{B.31}$$

That completes the proof.

Proof of (B.28): Consider any basic interaction $S_j = \{k_1, \ldots, k_{s_j}\}$, $j \in [J]$, then by Lemma B.2.1 we have that

$$\Omega_0 \cap \{B^{k_1} = \ldots = B^{k_{s_j}} = -1\} \subset \{S_j^- \subset \mathcal{F}\}. \tag{B.32}$$

Moreover, when $s_j = 1$, we also have that

$$\Omega_0 \cap \{B^{k_1} = +1\} \subset \{S_j^+ \subset \mathcal{F}\}. \tag{B.33}$$

Consequently, when $\tilde{S}$ is a union interaction as in Definition 9.1.1 it follows that

$$\Omega_0 \cap \{B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}}\} \subset \{\tilde{S}^{\pm} \subset \mathcal{F}\}, \tag{B.34}$$

which, shows (B.28).

Proof of (B.29): Assume that $\tilde{S}^{\pm}$ is not a union interaction. If any of the following is true:

- $\tilde{S}^{\pm}$ contains any noisy signed feature (k, b) that's not contained in $\cup_j S_j^+ \cup S_j^-$;
- for some signal feature $k \in \cup_j S_j$ we have that $(k, +1), (k, -1) \in \tilde{S}^{\pm}$;
- $|\tilde{S}^{\pm} \cap S_j^+| > 1$ for some $j \in [J]$;

Then by definition of $U(t)$ in (B.23), $P\left(\tilde{S}^{\pm} \subset \mathcal{F}\right) = 0$ and thus, (B.29) holds.

Thus, we can assume that $\tilde{S}^{\pm}$ contains no noisy features and there exists some interaction $j \in [J]$ with $s_j > 1$ such that $(S_j^- \cup S_j^+) \cap \tilde{S}^{\pm} \neq \emptyset$ and for some $(k, -1) \in S_j^-$ we have that $(k, -1) \notin \tilde{S}^{\pm}$.

First, assume that $(k,+1) \notin \tilde{S}^{\pm}$. Then, whenever t_{root} splits on feature k, we have that $\{\tilde{S}^{\pm} \subset \mathcal{F}\}$ implies[1] $B^k = -1$ and thus,

$$\begin{aligned}
P\left(\tilde{S}^{\pm} \subset \mathcal{F}|\mathcal{D}\right) &= \sum_{\tilde{k}\in[p]} P\left(\tilde{S}^{\pm} \subset \mathcal{F} \cap t_{\text{root}} \text{ splits on } \tilde{k}|\mathcal{D}\right) \\
&\leq \sum_{\tilde{k}\neq k} P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}} \cap t_{\text{root}} \text{ splits on } \tilde{k}|\mathcal{D}\right) \\
&+ P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}} \cap B^k = -1 \cap t_{\text{root}} \text{ splits on } k|\mathcal{D}\right) \\
&= \sum_{\tilde{k}\neq k} P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}}\right) P\left(t_{\text{root}} \text{ splits on } \tilde{k}|\mathcal{D}\right) \\
&+ P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}} \cap B^k = -1\right) P\left(t_{\text{root}} \text{ splits on } k|\mathcal{D}\right) \\
&= 0.5^{\tilde{s}}(1 - P(t_{\text{root}} \text{ splits on } k|\mathcal{D})) + 0.5^{\tilde{s}+1} P(t_{\text{root}} \text{ splits on } k|\mathcal{D}) \\
&\leq 0.5^{\tilde{s}} - 0.5^{\tilde{s}+1} P(t_{\text{root}} \text{ splits on } k|\mathcal{D}) \leq 0.5^{\tilde{s}}(1 - C_{\text{root}}/2),
\end{aligned}$$

where we made use of the fact that the tree T is independent of the Bernoulli random variables $\mathcal{B}$.

Second, assume that $(k,+1) \in \tilde{S}^{\pm}$. If $\tilde{S}^{\pm} \cap S_j^- \neq \emptyset$, then $\{\tilde{S}^{\pm} \subset \mathcal{F}\}$ implies[2] t_{root} does not split on k and thus

$$\begin{aligned}
P\left(\tilde{S}^{\pm} \subset \mathcal{F}\right) &\leq P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}}\right) P(t_{\text{root}} \text{ does not split on } k) \\
&= 0.5^{\tilde{s}} P(t_{\text{root}} \text{ does not split on } k) \leq 0.5^{\tilde{s}}(1 - C_{\text{root}}).
\end{aligned}$$

If $\tilde{S}^{\pm} \cap S_j^- = \emptyset$, let $k^\star \in S_j$ and $k^\star \neq k$. Because $(k,+1) \in \tilde{S}^{\pm}$, we can assume that $(k^\star,+1) \notin \tilde{S}^{\pm}$; otherwise $|\tilde{S}^{\pm} \cap S_j^+| > 1$, which implies $P\left(\tilde{S}^{\pm} \subset \mathcal{F}\right) = 0$. When t_{root} splits on $k^\star$, $\{\tilde{S}^{\pm} \subset \mathcal{F}\}$ implies[3] $B^{k^\star} = -1$ and thus,

$$\begin{aligned}
P\left(\tilde{S}^{\pm} \subset \mathcal{F}\right) &= \sum_{\tilde{k}\in[p]} P\left(\tilde{S}^{\pm} \subset \mathcal{F} \cap t_{\text{root}} \text{ splits on } \tilde{k}\right) \\
&\leq \sum_{\tilde{k}\neq k^\star} P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}} \cap t_{\text{root}} \text{ splits on } \tilde{k}\right) \\
&+ P\left(B^{k_1} = b_1 \cap \ldots \cap B^{k_{\tilde{s}}} = b_{\tilde{s}} \cap B^{k^\star} = -1 \cap t_{\text{root}} \text{ splits on } k^\star\right) \\
&\leq 0.5^{\tilde{s}}(1 - P(t_{\text{root}} \text{ splits on } k^\star)) + 0.5^{\tilde{s}+1} P(t_{\text{root}} \text{ splits on } k^\star) \leq 0.5^{\tilde{s}}(1 - C_{\text{root}}/2).
\end{aligned}$$

[1] Note that this requires the interactions to be disjoint, as otherwise the features in $\tilde{S}^{\pm} \cap (S_j^+ \cup S_j^-)$ may also appear in other interactions S_l with $l \neq j$ and $k \notin S_l$ and thus, even when $B^k = +1$ it is possible that $\tilde{S}^{\pm} \cap (S_j^+ \cup S_j^-) \subset \mathcal{F}$.

[2] Again, this requires the interactions to be disjoint, as otherwise the features in $\tilde{S}^{\pm} \cap (S_j^- \cup S_j^+) \setminus (k,+1)$ may also appear in other interactions S_l with $l \neq j$ and thus, even when t_{root} splits on k with $B^k = +1$, it is possible that $\tilde{S}^{\pm} \cap (S_j^- \cup S_j^+) \setminus (k,+1) \subset \mathcal{F}$.

[3] Again, this requires the interactions to be disjoint.

Thus, we have shown (B.29). □

Proof of the finite sample case

Filtering of desirable features and impurity

Recall that $R_{t,l} = R_t \cap \{X | X_{k_t} \leq \gamma_t\}$ and $R_{t,r} = R_t \cap \{X | X_{k_t} > \gamma_t\}$ denote the region corresponding to the left and right children for node t. In other words, node t divides the region R_t into $R_{t,l}$ and $R_{t,r}$. Recall that $N_n(t)$ is the number of samples in the region R_t, i.e., $N_n(t) = \sum_{i=1}^n \mathbf{1}(x_i \in R_t)$. We will use an equivalent formula for the impurity as in Lemma B.2.2.

Lemma B.2.2. *$\Delta_I^n(R_{t,l}, R_{t,r})$ defined in (9.4) in the main text is equivalent to (B.35):*

$$\Delta_I^n(R_{t,l}, R_{t,r}) = \frac{N_n(t_l)N_n(t_r)}{n(N_n(t_l) + N_n(t_r))} \left(\frac{1}{N_n(t_l)} \sum_{\mathbf{x}_i \in R_{t,l}} y_i - \frac{1}{N_n(t_r)} \sum_{\mathbf{x}_i \in R_{t,r}} y_i \right)^2. \tag{B.35}$$

Proof.

$$\begin{aligned}
\Delta_I^n(R_{t,l}, R_{t,r}) =& \frac{1}{n} \left(\sum_{\mathbf{x}_i \in R_t} (y_i - \frac{1}{N_n(t)} \sum_{\mathbf{x}_i \in R_t} y_i)^2 - \sum_{\mathbf{x}_i \in R_{t,l}} (y_i - \frac{1}{N_n(t_l)} \sum_{\mathbf{x}_i \in R_{t,l}} y_i)^2 - \sum_{\mathbf{x}_i \in R_{t,r}} (y_i - \frac{1}{N_n(t_r)} \sum_{\mathbf{x}_i \in R_{t,r}} y_i)^2 \right) \\
=& \frac{1}{n} \left(\sum_{\mathbf{x}_i \in R_t} y_i^2 - \frac{1}{N_n(t)} (\sum_{\mathbf{x}_i \in R_t} y_i)^2 - \sum_{\mathbf{x}_i \in R_{t,l}} y_i^2 + \frac{1}{N_n(t_l)} (\sum_{\mathbf{x}_i \in R_{t,l}} y_i)^2 - \sum_{\mathbf{x}_i \in R_{t,r}} y_i^2 + \frac{1}{N_n(t_r)} (\sum_{\mathbf{x}_i \in R_{t,r}} y_i)^2 \right) \\
=& \frac{1}{n} \left(-\frac{1}{N_n(t)} (\sum_{\mathbf{x}_i \in R_t} y_i)^2 + \frac{1}{N_n(t_l)} (\sum_{\mathbf{x}_i \in R_{t,l}} y_i)^2 + \frac{1}{N_n(t_r)} (\sum_{\mathbf{x}_i \in R_{t,r}} y_i)^2 \right).
\end{aligned}$$

If we denote $A = \sum_{\mathbf{x}_i \in R_{t,l}} y_i$ and $B = \sum_{\mathbf{x}_i \in R_{t,r}} y_i$, the above formula is the same as :

$$\begin{aligned}
& \frac{1}{n} \left(-\frac{1}{N_n(t)} (A+B)^2 + \frac{1}{N_n(t_l)} A^2 + \frac{1}{N_n(t_r)} B^2 \right) \\
=& \frac{1}{n} \left(\frac{N_n(t_r)}{N_n(t_l)N_n(t)} A^2 + \frac{N_n(t_l)}{N_n(t_r)N_n(t)} B^2 - \frac{2}{N_n(t)} AB \right) \\
=& \frac{N_n(t_l)N_n(t_r)}{nN_n(t)} \left(\frac{1}{N_n(t_l)^2} A^2 + \frac{1}{N_n(t_r)^2} B^2 - \frac{2}{N_n(t_l)N_n(t_r)} AB \right) = \frac{N_n(t_l)N_n(t_r)}{nN_n(t)} \left(\frac{1}{N_n(t_l)} A - \frac{1}{N_n(t_r)} B \right)^2.
\end{aligned}$$

□

Let $\mathscr{R}$ denote the set of axis-aligned hyper-rectangles obtained by splitting the unit hyper-rectangle consecutively, where each split satisfies A3. We study $\mathscr{R}$ because it contains all the possible rectangles that can represent region of a node in a tree. Let $\mathscr{R}_d$ be the set of rectangles obtained by splitting the unit hyper-rectangle d times, where each split satisfies assumption A3. Then $\mathscr{R} = \cup_{d \geq 1} \mathscr{R}_d$, and for any $R \in \mathscr{R}_d$, we have $\mu(R) \leq (1 - C_\gamma)^d$. For any region R, we denote N_R to be the number of points in R.

Lemma B.2.3. *Suppose that A3 is satisfied. Then for any $d \geq 1$,*

$$\max_{R\in\cup_{d_1>d}\mathscr{R}_{d_1}}\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R_1)-\mathbb{E}(Y\cdot\mathbf{1}(X\in R))\right|\leq C_Y\left(\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|\right)+2C_Y(1-C_\gamma)^d.$$

Proof of Lemma. For any $R_1 \in \mathscr{R}_{d_1}$, $d_1 > d$, there exists $R_0 \in \mathscr{R}_d$ such that $R_1 \subset R_0$. Therefore $N_{R_1} < N_{R_0}$, and

$$\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R_1)\right|\leq\frac{N_{R_1}}{n}C_Y<\frac{N_{R_0}}{n}C_Y$$

Since $R_0 \in \mathscr{R}_d$, we have

$$\frac{N_{R_0}}{n}\leq\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|+\max_{R\in\mathscr{R}_d}\mu(R)\leq\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|+(1-C_\gamma)^d. \tag{B.36}$$

Therefore

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R_1)-\mathbb{E}(Y\cdot\mathbf{1}(X\in R))\right|\\
\leq\ &\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R_1)\right|+|\mathbb{E}(Y\cdot\mathbf{1}(X\in R))|\\
\leq\ &\frac{N_{R_0}}{n}C_Y+C_Y(1-C_\gamma)^{d+1}\\
\leq\ &C_Y\left(\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|+(1-C_\gamma)^d\right)+C_Y(1-C_\gamma)^{d+1}\\
\leq\ &C_Y\left(\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|\right)+2C_Y(1-C_\gamma)^d.
\end{aligned}$$

Since R_1 is arbitrary, we have

$$\max_{R\in\cup_{d_1>d}\mathscr{R}_{d_1}}\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R_1)-\mathbb{E}(Y\cdot\mathbf{1}(X\in R))\right|\leq C_Y\left(\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right|\right)+2C_Y(1-C_\gamma)^d. \tag{B.37}$$

□

Proposition B.2.1. *Suppose that A1 and A3 hold true. Then*

$$\max_{R\in\mathscr{R}}\left|\frac{N_R}{n}-\mu(R)\right|\xrightarrow{p}0,$$

and

$$\max_{R\in\mathscr{R}}\left|\frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{1}(\mathbf{x}_i\in R)-\mathbb{E}(Y\cdot\mathbf{1}(X\in R))\right|\xrightarrow{p}0.$$

Proof. For any fixed d, let $G_n(\mathscr{R}_d)$ be the growth function for the set of rectangles $\mathscr{R}_d$ defined in Chapter 5.2 of Vapnik [93], i.e.,

$$G_n(\mathscr{R}_d) \triangleq \max_{\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}} \log \left| \left\{ (\mathbf{1}(y_1 \geq \theta, \mathbf{x}_1 \in R), \ldots, \mathbf{1}(y_n \geq \theta, \mathbf{x}_n \in R)) \middle| R \in \mathscr{R}_d, \theta \in \mathbb{R} \right\} \right|.$$

Here for any set A, $|A|$ denotes the number of elements in A.

We claim that $G_n(\mathscr{R}_d) \leq \log(n(2np)^d)$. This is because at each of d splits, we have at most p directions and at most n split points to choose from. Therefore, splitting d times can create no more than $(2np)^d$ different separations of the n data points. Furthermore, within each rectangle, the indicator functions $\mathbf{1}(y_i \geq \theta), \theta \in \mathbb{R}$ can at most create n separations.

Thus,

$$G_n(\cup_{d_0 \leq d} \mathscr{R}_{d_0}) \leq \log(d \exp(G_n(\mathscr{R}_d))) \leq \log(nd(2np)^d), \tag{B.38}$$

and

$$\frac{G_n(\cup_{d_0 \leq d} \mathscr{R}_{d_0})}{n} \leq \frac{\log(nd) + d\log(2n)}{n} + \frac{d \log p}{n} \to 0.$$

Therefore, by Theorem 5.1 of [93]:

Theorem 5.1 in [93]: Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ be a measurable set of bounded real-valued functions. Let G_n be the growth function of the indicator functions induced by Q, then we have the following inequality:

$$P\left\{ \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha) \right) > \epsilon \right\} \leq 4 \exp\left\{ \left(\frac{G_{2n}}{n} - \frac{(\epsilon - n^{-1})^2}{(B-A)^2} \right) n \right\}.$$

we have

$$\max_{R \in \cup_{d_0 \leq d} \mathscr{R}_{d_0}} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(Y \cdot \mathbf{1}(X \in R)) \right| \xrightarrow{p} 0. \tag{B.39}$$

Taking $Y = 1$ and we have

$$\max_{R \in \cup_{d_0 \leq d} \mathscr{R}_{d_0}} \left| \frac{N_R}{n} - \mu(R) \right| \xrightarrow{p} 0. \tag{B.40}$$

By Lemma B.2.3 and the above equation, we have

$$\begin{aligned} &\max_{R \in \cup_{d_1 > d} \mathscr{R}_{d_1}} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}(\mathbf{x}_i \in R_1) - \mathbb{E}(f(X) \cdot \mathbf{1}(X \in R_1)) \right| \\ \leq & C_Y \left(\max_{R \in \mathscr{R}_d} \left| \frac{N_R}{n} - \mu(R) \right| \right) + 2C_Y(1 - C_\gamma)^d \xrightarrow{p} 3C_Y(1 - C_\gamma)^d. \end{aligned} \tag{B.41}$$

Since that holds for any fixed $d > 0$, we know the left hand side of (B.41) converges to zero in probability. Combining (B.39) and (B.41), we have shown that:

$$\max_{R \in \mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(Y \cdot \mathbf{1}(X \in R)) \right| \xrightarrow{p} 0.$$

Since this holds for any bounded random variable Y, we can take $Y = 1$ and we have shown

$$\max_{R\in\mathscr{R}} \left| \frac{N_R}{n} - \mu(R) \right| \xrightarrow{p} 0. \tag{B.42}$$

That completes the proof. □

Proposition B.2.2 (Subgaussian case)**.** *Suppose that A3 holds true and* $(\log n)^{1+\delta} \log p/n \to 0$ *for some* $\delta > 0$*. Suppose* $Y = \mathbb{E}(Y|X) + Z$ *where* Z *is independent of* X *and* 1*-subgaussian. Then*

$$\max_{R\in\mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(Y \cdot \mathbf{1}(X \in R)) \right| \xrightarrow{p} 0.$$

Proof. Denote $f(X) = E(Y|X)$ and $C_Y = \sum_{j=0}^{J} |\beta_j|$. Then $|f(X)| \le C_Y$. Note that

$$\begin{aligned} & \max_{R\in\mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(Y \cdot \mathbf{1}(X \in R)) \right| \\ \le\ & \max_{R\in\mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i) \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(f(X) \cdot \mathbf{1}(X \in R)) \right| + \max_{R\in\mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} z_i \mathbf{1}(\mathbf{x}_i \in R) \right|. \end{aligned}$$

Here $z_i = y_i - f(\mathbf{x}_i)$ represents the noise terms. Our proof proceeds in the following two steps.

Step 1. Show that

$$\max_{R\in\mathscr{R}} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i) \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(f(X) \cdot \mathbf{1}(X \in R)) \right| \xrightarrow{p} 0. \tag{B.43}$$

Step 1 is similar to the proof of B.2.1 but the difference is that we need the convergence rate. Let

$$\delta_0 = \frac{\delta}{2\delta + 4},$$

and take

$$d = \left(\frac{n}{\log(np)} \right)^{\delta_0} \to \infty.$$

Let $G_n(\mathscr{R}_d)$ be the growth function for the set of rectangles $\mathscr{R}_d$. By (B.38), we have

$$\frac{G_n(\cup_{d_0 \le d} \mathscr{R}_{d_0})}{n} \le \frac{\log(nd) + d\log(2n)}{n} + \frac{d\log p}{n} = O\left(\frac{d\log(np)}{n} \right) = O\left(\left(\frac{\log(np)}{n} \right)^{1-\delta_0} \right) \to 0.$$

Therefore, by Theorem 5.1 of [93], we have

$$\max_{R\in\cup_{d_0 \le d} \mathscr{R}_{d_0}} \left| \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i) \mathbf{1}(\mathbf{x}_i \in R_1) - \mathbb{E}(f(X) \cdot \mathbf{1}(X \in R)) \right| = o\left(\left(\frac{\log(np)}{n} \right)^{1/2-\delta_0} \right) \xrightarrow{p} 0. \tag{B.44}$$

Since this holds for any bounded random variable Y, we can take $Y = 1$ and it follows that

$$\max_{R\in\mathscr{R}_d}\left|\frac{N_R}{n}-\mu(R)\right| \le \max_{R\in\cup_{d_0\le d}\mathscr{R}_{d_0}}\left|\frac{N_R}{n}-\mu(R)\right| = o\left(\left(\frac{\log(np)}{n}\right)^{1/2-\delta_0}\right) \xrightarrow{p} 0. \tag{B.45}$$

Since $d\to\infty$, $(1-C_\gamma)^d\to 0$. Therefore, by Lemma B.2.3, we have

$$\max_{R\in\cup_{d_1>d}\mathscr{R}_{d_1}}\left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\mathbf{1}(\mathbf{x}_i\in R_1)-\mathbb{E}(f(X)\cdot\mathbf{1}(X\in R))\right| \xrightarrow{p} 0. \tag{B.46}$$

Combining (B.44) and (B.46), (B.43) is proved.

Step 2. Show that

$$\max_{R\in\mathscr{R}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right| \xrightarrow{p} 0 \tag{B.47}$$

Note that

$$\max_{R\in\mathscr{R}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right| = \max\left\{\max_{R\in\cup_{d_0\le d}\mathscr{R}_{d_0}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|, \max_{R\in R\in\cup_{d_1>d}\mathscr{R}_{d_1}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|\right\}.$$

Therefore, it suffices to prove that both of the two terms on the right hand size converges to 0 in probability. We begin with the first term: $\max_{R\in\cup_{d_0\le d}\mathscr{R}_{d_0}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|$. Since X and Z are independent and Z is 1-subgaussian, by Hoeffding inequality,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|\ge\epsilon/2\;\middle|\;X\right) = P\left(\left|\frac{1}{n}\sum_{i=1}^{N_R} z_i\right|\ge\epsilon/2\right) \le 2\exp\left(-\frac{n^2\epsilon^2}{8N_R}\right).$$

for any rectangle R. Therefore by union bound,

$$\begin{aligned}&P\left(\max_{R\in\cup_{d_0\le d}\mathscr{R}_{d_0}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|\ge\epsilon/2\;\middle|\;X\right)\\ \le& 2\exp(G_n(\cup_{d_0\le d}\mathscr{R}_{d_0}))\exp\left(-\frac{n\epsilon^2}{8}\right) \le 2\exp\left(\log(nd(2np)^d)-\frac{n\epsilon^2}{8}\right)\to 0.\end{aligned}$$

for any $\epsilon>0$. Since the above upper bound on the probability is independent of X, we conclude that

$$\max_{R\in\cup_{d_0\le d}\mathscr{R}_{d_0}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right| \xrightarrow{p} 0.$$

We now turn to the second term $\max_{R\in R\in\cup_{d_1>d}\mathscr{R}_{d_1}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbf{1}(\mathbf{x}_i\in R)\right|$. Let $\mathscr{R}^{s_0}$ be the set of rectangles with at most $s_0 = n/(\log n)^{1/2+\delta_0}$ samples, then $\log|\mathscr{R}^{s_0}|\le(s_0+1)\log n$. By

union bound,

$$P\left(\max_{R\in\mathscr{R}^{s_0}}\left|\frac{1}{n}\sum_{i=1}^{n}z_i\mathbf{1}(\mathbf{x}_i\in R)\right|\geq\epsilon/2\;\middle|\;X\right)$$
$$\leq 2\exp(\log|\mathscr{R}^{s_0}|)\exp\left(-\frac{n\epsilon^2}{8}\right)\leq 2\exp\left((s_0+1)\log n-\frac{n^2\epsilon^2}{8s_0}\right)\to 0.$$

Therefore,

$$\max_{R\in\mathscr{R}^{s_0}}\left|\frac{1}{n}\sum_{i=1}^{n}z_i\mathbf{1}(\mathbf{x}_i\in R)\right|\to 0.$$

Hence, to prove (B.47), it suffices to show that $\cup_{d_1>d}\mathscr{R}_{d_1}\subset\mathscr{R}^{s_0}$ with probability tending to 1. Note that by definition of δ_0, $\frac{1/2+\delta_0}{1/2-\delta_0}=1+\delta$. Therefore

$$\left(\frac{\log(np)}{n}\right)^{\frac{1}{2}-\delta_0}(\log n)^{\frac{1}{2}+\delta_0}=\left(\frac{\log(np)(\log n)^{1+\delta}}{n}\right)^{\frac{1}{2}-\delta_0}=\left(\frac{(\log n)^{2+\delta}+\log p(\log n)^{1+\delta}}{n}\right)^{\frac{1}{2}-\delta_0}\to 0.$$

By (B.36) and (B.45), we have

$$\max_{R\in\cup_{d_1>d}\mathscr{R}_{d_1}}N_R\leq\max_{R\in\mathscr{R}_d}N_R=o\left(n\left(\frac{\log(np)}{n}\right)^{1/2-\delta_0}\right)=o(s_0).$$

Therefore, $\max_{R\in\cup_{d_1>d}\mathscr{R}_{d_1}}N_R\leq s_0$ with probability tending to 1. The proof is now complete. □

Define population impurity decrease $\Delta_I(t)$ at a node t to be

$$\Delta_I(t)=\mathrm{Var}(Y|R_t)-\frac{\mu(R_{t_l})}{\mu(R_t)}\mathrm{Var}(Y|R_{t_l})-\frac{\mu(R_{t_r})}{\mu(R_t)}\mathrm{Var}(Y|R_{t_r}).\tag{B.48}$$

Similar to Lemma B.2.2, we know it is equivalent to:

$$\Delta_I(R_{t,l}(\gamma;k),R_{t,r}(\gamma;k))=\frac{\mu(R_{t,l}(\gamma;k))\mu(R_{t,r}(\gamma;k))}{\mu(R_t(\gamma;k))}\Big[\mathbb{E}(Y|X\in R_{t,l}(\gamma;k))-\mathbb{E}(Y|X\in R_{t,r}(\gamma;k))\Big]^2.\tag{B.49}$$

The following proposition shows that the finite-sample impurity decrease converges to the population impurity decrease uniformly.

Proposition B.2.3. *With A1 and A3, we have the following two uniform convergence results:*

a. $\max_{R\in\mathscr{R}}\left|\frac{N_R}{n}-\mu(R)\right|\xrightarrow{p}0$

b. $\sup_{R_{t,l},R_{t,r}\in\mathscr{R}}\left|\Delta_I^n(R_{t,l},R_{t,r})-\Delta_I(R_{t,l},R_{t,r})\right|\xrightarrow{p}0.$

Proof. **a.** That's the direct conclusion of Proposition B.2.1.

b. Let $f(x_1, x_2, y_1, y_2) = \frac{x_1 x_2}{x_1+x_2}(y_1 - y_2)^2$. Then f is a Lipschitz function on $[0,1] \times [0,1] \times [-C_Y - 1, C_Y + 1] \times [-C_Y - 1, C_Y + 1]$. Use the fact that $\max_{R\in\mathscr{R}} \left| \frac{1}{n}\sum_{i=1}^n y_i \mathbf{1}(\mathbf{x}_i \in R) - \mathbb{E}(Y \cdot \mathbf{1}(X \in R)) \right| \xrightarrow{p} 0$ in proposition B.2.1 and the fact $\max_{R\in\mathscr{R}} \left| \frac{N_R}{n} - \mu(R) \right| \xrightarrow{p} 0$ in a), by continuous mapping theorem, we have

$$\sup_{R_{t,l}, R_{t,r} \in \mathscr{R}} \left| \Delta_I^n(R_{t,l}, R_{t,r}) - \Delta_I(R_{t,l}, R_{t,r}) \right| \xrightarrow{p} 0.$$

□

Now we analyze the impurity decrease at each node of a tree. We consider three families of trees:$\mathcal{T}_0$, $\mathcal{T}_1$ and $\mathcal{T}_2$:

$$\mathcal{T}_0 \triangleq \{\text{Any tree that satisfies A3}\}.$$

$$\mathcal{T}_1 \triangleq \{\text{Any CART tree that satisfies A3 and A5}\}.$$
$$\mathcal{T}_2 \triangleq \{\text{Any CART tree that satisfies A3, A5, and A4}\}.$$

$\mathcal{T}_1$ is the family of CART trees that satisfy our assumptions but m_{try} can be arbitrary. $\mathcal{T}_1$ is more restricted than $\mathcal{T}_0$ in the sense that the threshold γ_t of any node t of any tree in $\mathcal{T}_1$ must maximize the finite sample impurity decrease (9.4). Thus, $\mathcal{T}_1$ depends on the data. For any $T \in \mathcal{T}_0$ and any $t \in T$ such that $\mathring{U}(t) \neq \emptyset$, its region R_t is a rectangle:

$$R_t = \{x \in \mathbb{R}^p | \forall \ell \in [p], c_{low,\ell} < x_\ell \leq c_{high,\ell}\}. \tag{B.50}$$

where $c_{low,\ell}, c_{high,\ell} \in [0,1]$.

By the definition of desirable feature set $U(t)$, we have its equivalent formula:

$$U(t) \triangleq \cup_{j\in[J]: S_j^+ \cap \mathrm{F}^{\pm}(t) = \emptyset} S_j / \mathrm{F}(t).$$

Define the set of noisy features to be its complement: $[p]/U(t)$. We can also define

$$\mathring{U}(t) \triangleq \cup_{j\in[J]: S_j^+ \cap \dot{\mathrm{F}}^{\pm}(t) = \emptyset} S_j / \mathrm{F}(t).$$

Since $\mathrm{F}^{\pm}(t) \subset \dot{\mathrm{F}}^{\pm}(t)$, $\mathring{U}(t) \subset U(t)$. For any γ, denote $R_{t,l}(\gamma; k) = R_t \cap \{X | X_k \leq \gamma\}$ and $R_{t,r}(\gamma; k) = R_t \cap \{X | X_k > \gamma\}$. First, for any node $t \in T$ and any $k \in \mathring{U}(t)$, we have a characterization for the impurity decrease:

Lemma B.2.4. *For any $T \in \mathcal{T}_0$, $t \in T$, $j \in [J]$, $k \in S_j \cap U(t)$, and $\gamma \in (0,1)$,*

$$\begin{aligned}
&\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \\
=&\mu(R_t) \cdot \beta_j^2 P(\forall\, \ell \in S_j/\{k\},\ X_\ell \leq \gamma_\ell | X \in R_t)^2 \cdot \\
&\Big(\mathbf{1}(\gamma \leq \gamma_k) \cdot \frac{(1-\gamma_k)^2 \gamma}{(1-\gamma)} + \mathbf{1}(\gamma > \gamma_k) \cdot \frac{\gamma_k^2 (1-\gamma)}{\gamma} \Big).
\end{aligned}$$

Proof of Lemma B.2.4. Since $k \in U(t)$, we know that k is not in $\mathrm{F}(t)$. That means any of t's parents do not split on k. In other words, R_t does not have any constraints for feature k, i.e., $c_{low,k} = 0$ and $c_{high,k} = 1$. Thus, we know that

$$\mu(R_{t,l}(\gamma; k)) = \mu(R_t) \cdot \gamma \tag{B.51}$$

and

$$\mu(R_{t,r}(\gamma; k)) = \mu(R_t) \cdot (1 - \gamma). \tag{B.52}$$

Recall that Δ_I in (B.48) has its equivalent formula (B.49):

$$\begin{aligned}&\Delta_I(R_{t,l}(\gamma; k), R_{t,r}(\gamma; k)) = \\ &\frac{\mu(R_{t,l}(\gamma; k))\mu(R_{t,r}(\gamma; k))}{\mu(R_t(\gamma; k))}\Big[\mathbb{E}(Y|X \in R_{t,l}(\gamma; k)) - \mathbb{E}(Y|X \in R_{t,r}(\gamma; k))\Big]^2\end{aligned}$$

where the conditional expectations are

$$\mathbb{E}(Y|X \in R_{t,l}(\gamma; k)) = \sum_{j'=1}^{J} \beta_{j'} P\left(\forall \ell \in S_{j'},\ X_\ell \le \gamma_\ell \middle| X \in R_{t,l}(\gamma; k)\right), \tag{B.53}$$

and

$$\mathbb{E}(Y|X \in R_{t,r}(\gamma; k)) = \sum_{j'=1}^{J} \beta_{j'} P\left(\forall \ell \in S_{j'},\ X_\ell \le \gamma_\ell \middle| X \in R_{t,r}(\gamma; k)\right). \tag{B.54}$$

Now we will analyze (B.53) and (B.54). To ease the notations, let's define the following three events:

$$A_{j'} = \{X_\ell \le \gamma_\ell,\ \forall \ell \in S_{j'}\}, \tag{B.55}$$
$$B = \{X \in R_t\}, \tag{B.56}$$
$$C_k = \{X_k \le \gamma\}. \tag{B.57}$$

Then (B.53) becomes $\sum_{j'=1}^{J} \beta_{j'} P(A_{j'}|BC_k)$. Because R_t has not constraints on k, B does not involve feature k. When $j' \neq j$ (namely, $k \notin S_{j'}$), $A_{j'}$ also does not involve feature k. Thus, C_k is independent of $(A_{j'}, B)$, which implies $P(A_{j'}|BC_k) = \frac{P(A_{j'}BC_k)}{P(BC_k)} = \frac{P(A_{j'}B)P(C_k)}{P(B)P(C_k)} = P(A_{j'}|B)$. Similarly this holds for (B.54). Therefore, when $j' \neq j$:

$$P\left(\forall \ell \in S_{j'},\ X_\ell \le \gamma_\ell \middle| X \in R_{t,l}(\gamma; k)\right) = P\left(\forall \ell \in S_{j'},\ X_\ell \le \gamma_\ell \middle| X \in R_{t,r}(\gamma; k)\right).$$

When $j' = j$,

$$\begin{aligned}
&P\left(\forall \ell \in S_j,\ X_\ell \le \gamma_\ell \middle| X \in R_{t,l}(\gamma;k)\right) - P\left(\forall \ell \in S_j,\ X_\ell \le \gamma_\ell \middle| X \in R_{t,r}(\gamma;k)\right) \\
=&P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t)\cdot \\
&\qquad \Big(P(X_k \le \gamma_k | X_k \le \gamma) - P(X_k \le \gamma_k | X_k > \gamma)\Big) \\
=&P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t)\cdot \\
&\qquad \Big(\mathbf{1}(\gamma \le \gamma_k)\cdot \frac{1-\gamma_k}{1-\gamma} + \mathbf{1}(\gamma > \gamma_k)\cdot \frac{\gamma_k}{\gamma}\Big).
\end{aligned}$$

Therefore, (B.49) becomes:

$$\begin{aligned}
&\frac{\mu(R_{t,l}(\gamma;k))\mu(R_{t,r}(\gamma;k))}{\mu(R_t)}\Big(\mathbb{E}(Y|X \in R_{t,l}(\gamma;k)) - \mathbb{E}(Y|X \in R_{t,r}(\gamma;k))\Big)^2 \\
&= \mu(R_t)\gamma(1-\gamma)\cdot \beta_j^2 P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t)^2 \cdot \\
&\qquad \Big(\mathbf{1}(\gamma \le \gamma_k)\cdot \frac{(1-\gamma_k)^2}{(1-\gamma)^2} + \mathbf{1}(\gamma > \gamma_k)\cdot \frac{\gamma_k^2}{\gamma^2}\Big) \\
&= \mu(R_t)\cdot \beta_j^2 P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t)^2 \cdot \\
&\qquad \Big(\mathbf{1}(\gamma \le \gamma_k)\cdot \frac{(1-\gamma_k)^2\gamma}{(1-\gamma)} + \mathbf{1}(\gamma > \gamma_k)\cdot \frac{\gamma_k^2(1-\gamma)}{\gamma}\Big)
\end{aligned}$$

That completes the proof. □

Lemma B.2.5. *For $T \in \mathcal{T}_0$, $t \in T$, if there exists $j \in [J]$ and $k \in S_j$ such that $k \in \mathring{U}(t)$, then*

$$P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t) \ge C_\gamma^{s_j - 1}.$$

Proof of Lemma B.2.5. Because $k \in S_j$ and $k \in \mathring{U}(t)$, we know that $S_j^+ \cap \dot{\mathrm{F}}^{\pm}(t) = \emptyset$. That means node t is not at the right branch of any node that splits on features in S_j. Thus,

$$c_{low,\ell} = 0 \text{ when } \ell \in S_j. \tag{B.58}$$

Also, $c_{high,k} = 1$ and $c_{low,k} = 0$ because $k \in \mathring{U}(t)$. Then, $P(\forall\ \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t)$ is

$$\begin{aligned}
&\frac{P(\forall \ell \in S_j/\{k\}\ X_\ell \le \gamma_\ell, X \in R_t)}{\mu(R_t)} \\
(\text{Due to (B.58)}) =&\frac{\prod_{\ell \in [p]/S_j}(c_{high,\ell} - c_{low,\ell}) \prod_{\ell \in S_j/\{k\}} \min(c_{high,\ell}, \gamma_\ell)}{\mu(R_t)} \\
\ge&\frac{\prod_{\ell \in [p]/S_j}(c_{high,\ell} - c_{low,\ell}) \prod_{\ell \in S_j/\{k\}} c_{high,\ell}\cdot \gamma_\ell}{\mu(R_t)} \\
=&\frac{\mu(R_t)\cdot \prod_{\ell \in S_j/\{k\}} \gamma_\ell}{\mu(R_t)} \\
\ge&C_\gamma^{s_j - 1}.
\end{aligned}$$

That completes the proof. □

Lemma B.2.6. *Assume A1 holds. For any fixed $\epsilon > 0$,*

$$P\left(\inf_{T\in\mathcal{T}_0} \min_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \min_{k\in\mathring{U}(t)} \sup_{\gamma\in[C_\gamma,1-C_\gamma]} \Delta_I^n(R_{t,l}(\gamma;k),R_{t,r}(\gamma;k)) > \frac{\epsilon}{4}C_\beta^2 C_\gamma^{2\max_j s_j-1}\right)\to 1. \tag{B.59}$$

Proof. First of all, we know from Proposition B.2.3 that $\sup_{R_t\in\mathscr{R}}|\Delta_I^n(R_t)-\Delta_I(R_t)|\xrightarrow{p}0$. Thus, in order to prove (B.59), we only need to show that

$$\inf_{T\in\mathcal{T}_0} \min_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \min_{k\in\mathring{U}(t)} \Delta_I(R_{t,l}(\gamma_k;k),R_{t,r}(\gamma_k;k)) > \frac{\epsilon}{2}C_\beta^2 C_\gamma^{2\max_j s_j-1}. \tag{B.60}$$

Recall that γ_k is ground-truth threshold of feature k in the interaction. Here we can drop $\max_{\gamma\in[C_\gamma,1-C_\gamma]}$ and use γ_k because that results in a lower bound of the previous equation. Based on Lemma B.2.4, we know that

$$\begin{aligned}
&\Delta_I(R_{t,l}(\gamma_k;k),R_{t,r}(\gamma_k;k))\\
=&\mu(R_t)\cdot\beta_j^2 P(\forall\,\ell\in S_j/\{k\},\ X_\ell\leq\gamma_\ell|X\in R_t)^2\cdot(1-\gamma_k)\gamma_k\\
\geq&\frac{1}{2}C_\gamma C_\beta^2\epsilon\cdot P(\forall\,\ell\in S_j/\{k\},\ X_\ell\leq\gamma_\ell|X\in R_t)^2.
\end{aligned}$$

The second inequality is due to $\mu(R_t)\geq\epsilon$, $\gamma_k(1-\gamma_k)\geq C_\gamma(1-C_\gamma)\geq\frac{1}{2}C_\gamma$ and $\beta_j\geq C_\beta$. Then using Lemma B.2.5 leads to the conclusion. □

For a node t, denote $\gamma_{t,k}^*=\operatorname{argmax}_{\gamma\in[C_\gamma,1-C_\gamma]}\Delta_I^n(R_{t,l}(\gamma;k),R_{t,r}(\gamma;k))$.

Lemma B.2.7. *Suppose A1 holds true, we have*

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} |\gamma_{t,k}^*-\gamma_k|\xrightarrow{p}0.$$

Proof. To simplify the notation in the proof, let us denote

$$a_n=\Delta_I^n(R_{t,l}(\gamma_{t,k}^*;k),R_{t,r}(\gamma_{t,k}^*;k)),\quad a=\Delta_I(R_{t,l}(\gamma_{t,k}^*;k),R_{t,r}(\gamma_{t,k}^*;k)),$$

and

$$b_n=\Delta_I^n(R_{t,l}(\gamma_k;k),R_{t,r}(\gamma_k;k)),\quad b=\Delta_I(R_{t,l}(\gamma_k;k),R_{t,r}(\gamma_k;k)).$$

Using Proposition B.2.3, we have

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} \left|a_n - a\right| \xrightarrow{p} 0. \tag{B.61}$$

By Lemma B.2.6 (see (B.60)), we know the second term is bounded uniformly above zero:

$$\inf_{T\in\mathcal{T}_0} \min_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \min_{k\in\mathring{U}(t)} a \geq \frac{\epsilon}{2} C_\beta^2 C_\gamma^{2\max_j s_j - 1}.$$

Thus, the ratio converges to 1 in probability:

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} \left|\frac{a_n}{a} - 1\right| \xrightarrow{p} 0. \tag{B.62}$$

Similarly, this applies to b_n and b, too.

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} \left|\frac{b_n}{b} - 1\right| \xrightarrow{p} 0. \tag{B.63}$$

So by the continuous mapping theorem, we know that

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} \left|\frac{b_n}{a_n}\frac{a}{b} - 1\right| \xrightarrow{p} 0$$

Because $\gamma_{t,k}^*$ maximizes Δ_I^n and γ_k maximizes Δ_I, $a_n \geq b_n$ and $a \leq b$. Thus $\frac{b_n}{a_n}\frac{a}{b} \leq \frac{a}{b} \leq 1$. Therefore, we know

$$\sup_{T\in\mathcal{T}_0} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon,\\ \mathring{U}(t)\neq\emptyset}} \max_{k\in\mathring{U}(t)} 1 - \frac{a}{b} \xrightarrow{p} 0.$$

By Lemma B.2.4, we know that

$$\begin{aligned} a =& \mu(R_t)\cdot\beta_j^2 P(\forall\ \ell\in S_j/\{k\},\ X_\ell \leq \gamma_\ell | X\in R_t)^2 \cdot \\ & \Big(\mathbf{1}(\gamma_{t,k}^* \leq \gamma_k)\cdot\frac{(1-\gamma_k)^2\gamma_{t,k}^*}{(1-\gamma_{t,k}^*)} + \mathbf{1}(\gamma_{t,k}^* > \gamma_k)\cdot\frac{\gamma_k^2(1-\gamma_{t,k}^*)}{\gamma_{t,k}^*}\Big), \\ b =& \mu(R_t)\cdot\beta_j^2 P(\forall\ \ell\in S_j/\{k\},\ X_\ell \leq \gamma_\ell | X\in R_t)^2\cdot\gamma_k(1-\gamma_k) \end{aligned}$$

Thus the ratio is

$$\frac{a}{b} = \mathbf{1}(\gamma^*_{t,k} \le \gamma_k) \cdot \frac{(1-\gamma_k)\gamma^*_{t,k}}{\gamma_k(1-\gamma^*_{t,k})} + \mathbf{1}(\gamma^*_{t,k} > \gamma_k) \cdot \frac{\gamma_k(1-\gamma^*_{t,k})}{(1-\gamma_k)\gamma^*_{t,k}}.$$

When $\gamma^*_{t,k} \le \gamma_k$,

$$\begin{aligned} 1 - \frac{a}{b} &= 1 - \frac{(1-\gamma_k)\gamma^*_{t,k}}{\gamma_k(1-\gamma^*_{t,k})} \\ &= \frac{\gamma_k - \gamma^*_{t,k}}{\gamma_k(1-\gamma^*_{t,k})} \ge \gamma_k - \gamma^*_{t,k}. \end{aligned}$$

Similarly, when $\gamma^*_{t,k} \ge \gamma_k$, $1 - \frac{a}{b} \ge \gamma^*_{t,k} - \gamma_k$. Thus, $1 - a/b \ge |\gamma_k - \gamma^*_{t,k}| \ge 0$. Thus, by the Squeeze theorem, we complete the proof. □

Lemma B.2.8. *With A1, the following statements are true:*

i) For any fixed $\epsilon, \delta > 0$,

$$P\left(\inf_{T \in \mathcal{T}_1(\mathcal{D})} \min_{\substack{t \in T, \mu(R_t) \ge \epsilon \\ U(t) \ne \emptyset}} \min_{j \in [J]} \min_{k \in S_j \cap U(t)} P(\forall\, \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t; \mathcal{D}) - C_\gamma^{s_j - 1} \ge -\delta \right) \to 1.$$

ii) for any fixed $\epsilon > 0$,

$$P\left(\inf_{T \in \mathcal{T}_1(\mathcal{D})} \min_{\substack{t \in T, \mu(R_t) \ge \epsilon, \\ U(t) \ne \emptyset}} \min_{k \in U(t)} \sup_{\gamma \in [C_\gamma, 1-C_\gamma]} \Delta^n_I(R_{t,l}(\gamma; k), R_{t,r}(\gamma; k)) > \frac{\epsilon}{4} C_\beta^2 C_\gamma^{2 \max_j s_j - 1} \right) \to 1. \tag{B.64}$$

iii)

$$\sup_{T \in \mathcal{T}_1(\mathcal{D})} \max_{\substack{t \in T, \mu(R_t) \ge \epsilon, \\ U(t) \ne \emptyset}} \max_{k \in U(t)} |\gamma^*_{t,k} - \gamma_k| \xrightarrow{p} 0.$$

Proof. We use math induction to show that the above statements hold for any $L \ge 0$:

i) For any fixed $\epsilon, \delta > 0$,

$$P\left(\inf_{T \in \mathcal{T}_1(\mathcal{D})} \min_{\substack{t \in T, \mu(R_t) \ge \epsilon, U(t) \ne \emptyset, \\ \sum_{j=1}^J |\dot{\mathrm{F}}^{\pm}(t) \cap S_j^+| \le L}} \min_{j \in [J]} \min_{k \in S_j \cap U(t)} P(\forall\, \ell \in S_j/\{k\},\ X_\ell \le \gamma_\ell | X \in R_t; \mathcal{D}) - C_\gamma^{s_j - 1} \ge -\delta \right) \to 1.$$

ii) for any fixed $\epsilon > 0$,

$$P\left(\inf_{T \in \mathcal{T}_1(\mathcal{D})} \min_{\substack{t \in T, \mu(R_t) \ge \epsilon, U(t) \ne \emptyset, \\ \sum_{j=1}^J |\dot{\mathrm{F}}^{\pm}(t) \cap S_j^+| \le L}} \min_{k \in U(t)} \sup_{\gamma \in [C_\gamma, 1-C_\gamma]} \Delta^n_I(R_{t,l}(\gamma; k), R_{t,r}(\gamma; k)) > \frac{\epsilon}{4} C_\beta^2 C_\gamma^{2 \max_j s_j - 1} \right) \to 1. \tag{B.65}$$

iii)

$$\sup_{T\in\mathcal{T}_1(\mathcal{D})} \max_{\substack{t\in T, \mu(R_t)\geq\epsilon, U(t)\neq\emptyset,\\ \sum_{j=1}^J |\dot{\mathrm{F}}^{\pm}(t)\cap S_j^+|\leq L}} \max_{k\in U(t)} |\gamma^*_{t,k}-\gamma_k| \xrightarrow{p} 0.$$

If those statements are true, then our proof is complete because for any node t, $\sum_{j=1}^J |\dot{\mathrm{F}}^{\pm}(t)\cap S_j^+| \leq \sum_j s_j$, which is a constant.

When $L=0$, $U(t)\neq\emptyset$ and $\sum_j |\dot{\mathrm{F}}^{\pm}(t)\cap S_j^+| = 0$ implies that $U(t) = \cup_{j=1}^J S_j/F(t) = \mathring{U}(t)\neq\emptyset$. Then the statement holds because of Lemmas B.2.5, B.2.6, and B.2.7.

Suppose the statement holds for $L=L_0$, and let us consider the case $L=L_0+1$:

i): For $k\in S_j\cap U(t)$, we know that $S_j^+\cap \mathrm{F}^{\pm}(t)=\emptyset$. Now consider $S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)$: if it is also empty, then $k\in\mathring{U}(t)$, then i) holds because of Lemma B.2.5. Let's consider the case when $S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)\neq\emptyset$. For any $\ell\in S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)$, some parent nodes of t are split on feature ℓ and node t is at the left branch of the first parent node that is split on ℓ. In other words, this is the scenario where $(\ell,-1)$ first appears in the path and then $(\ell,+1)$ appears later. Denote that first parent node that is split on ℓ to be $t_{parent,\ell}$. Since none of $t_{parent,\ell}$'s parent nodes are split on ℓ, $\ell\in S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)$ but not $S_j^+\cap\dot{\mathrm{F}}^{\pm}(t_{parent,\ell})$. Since $\dot{\mathrm{F}}^{\pm}(t_{parent,\ell})$ is a subset of $\dot{\mathrm{F}}^{\pm}(t)$, we know that $\sum_{j=1}^J |S_j^+\cap\dot{\mathrm{F}}^{\pm}(t_{parent,\ell})|\leq L_0$. Also, because $S_j^+\cap\dot{\mathrm{F}}^{\pm}(t_{parent,\ell})=\emptyset$ and $\ell\notin\dot{F}(t_{parent,\ell})$, $\ell\in U(t_{parent,\ell})$. Then by the induction condition iii), we know that $\gamma^*_{t_{parent,\ell},\ell}\xrightarrow{p}\gamma_\ell$. Because t is at the left branch of $t_{parent,\ell}$, the upper bound in R_t for feature ℓ, i.e., $c_{high,\ell}$, is smaller or equal to $\gamma^*_{t_{parent,\ell},\ell}$. In other words, for any fixed $\delta>0$, we know that

$$P\left(\sup_{T\in\mathcal{T}_1(\mathcal{D})} \max_{\substack{t\in T,\mu(R_t)\geq\epsilon,U(t)\neq\emptyset,\\ \sum_{j=1}^J|\dot{\mathrm{F}}^{\pm}(t)\cap S_j^+|\leq L_0+1}} \max_{\substack{j\in[J]\\ S_j^+\cap\mathrm{F}^{\pm}(t)=\emptyset}} \max_{(\ell,+1)\in S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)} c_{high,\ell}-\gamma_\ell>\delta\right)\xrightarrow{p}0.$$

For any l such that $\ell\in S_j$ but $(\ell,+1)\notin S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)$, $c_{low,\ell}=0$. Note that $c_{high,k}=1$ and $c_{low,k}=0$ because $k\in U(t)$. Then, $P(\forall\,\ell\in S_j/\{k\},\ X_\ell\leq\gamma_\ell|X\in R_t;\mathcal{D})$ is

$$\begin{aligned}
&\frac{P(\forall\ell\in S_j/\{k\}\ X_\ell\leq\gamma_\ell, X\in R_t;\mathcal{D})}{\mu(R_t)}\\
=&\frac{\prod_{\ell\in[p]/S_j}(c_{high,\ell}-c_{low,\ell})\prod_{\ell\in S_j/\{k\}}\max(\min(c_{high,\ell},\gamma_\ell)-c_{low,\ell},0)}{\mu(R_t)}\\
=&\frac{\prod_{\ell\in[p]/S_j}(c_{high,\ell}-c_{low,\ell})\prod_{(\ell,+1)\in S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)}(c_{high,\ell}-c_{low,\ell}+o_p(1))\prod_{\ell\in S_j/\{k\},(\ell,+1)\notin S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)}\min(c_{high,\ell},\gamma_\ell)}{\mu(R_t)}\\
\geq&\frac{\prod_{\ell\in[p]/S_j}(c_{high,\ell}-c_{low,\ell})\prod_{(\ell,+1)\in S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)}(c_{high,\ell}-c_{low,\ell})\prod_{\ell\in S_j/\{k\},(\ell,+1)\notin S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)}c_{high,\ell}\cdot\gamma_\ell}{\mu(R_t)}+o_p(1)\\
\geq&\frac{\mu(R_t)\cdot\prod_{\ell\in S_j/\{k\},(\ell,+1)\notin S_j^+\cap\dot{\mathrm{F}}^{\pm}(t)}\gamma_\ell}{\mu(R_t)}+o_p(1)\\
\geq&C_\gamma^{s_j-1}+o_p(1).
\end{aligned}$$

That completes the proof for i).

ii): Given i), ii) should be obvious following the same proof in Lemma B.2.6.

iii) Given ii), iii) should follow the same proof in Lemma B.2.7.

Thus, we have finished the math induction and proved the statements. □

Lemma B.2.9. *For any tree $T \in \mathcal{T}_1$ and any node $t \in T$, the noisy features correspond to a nearly zero impurity decrease, i.e.*

$$\sup_{T\in\mathcal{T}_1} \max_{t\in T} \max_{k\in[p]/U(t)} \sup_{\gamma\in[0,1]} \Delta_I^n(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \xrightarrow{p} 0. \tag{B.66}$$

Proof. By Proposition B.2.3, we only need to show that

$$\sup_{T\in\mathcal{T}_1} \max_{t\in T} \max_{k\in[p]/U(t)} \sup_{\gamma\in[0,1]} \Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \xrightarrow{p} 0. \tag{B.67}$$

For $k \in [p]/U(t)$, either $k \in [p]/\bigcup_{j=1}^J S_j$ or $k \in \bigcup_{j=1}^J S_j/U(t)$. We will analyze these two cases separately.

a) When $k \in [p]/\bigcup_{j=1}^J S_j$, for any $j' \in [J]$, k is not contained in $S_{j'}$. Because different features are independent, X_k is independent from $X \in \{X|\forall\ \ell \in S_{j'},\ X_\ell \le \gamma_\ell\}$. Therefore, for any $j' \in [J]$, we have $P(\forall\ \ell \in S_{j'},\ X_\ell \le \gamma_\ell | X \in R_{t,l}(\gamma;k)) = P(\forall\ \ell \in S_{j'},\ X_\ell \le \gamma_\ell | X \in R_{t,r}(\gamma;k))$. That implies $\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) = 0$.

b) When there exists j such that $k \in S_j/U(t)$. For $j' \neq j$, by a similar deduction as in a), we know that $P(\forall\ \ell \in S_{j'},\ X_\ell \le \gamma_\ell | X \in R_{t,l}(\gamma;k)) = P(\forall\ \ell \in S_{j'},\ X_\ell \le \gamma_\ell | X \in R_{t,r}(\gamma;k))$. The impurity decrease $\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k))$ becomes

$$\frac{\mu(R_{t,l}(\gamma;k))\mu(R_{t,r}(\gamma;k))}{\mu(R_t)}\beta_j^2 \cdot \tag{B.68}$$
$$\Big(P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell | X \in R_{t,l}(\gamma;k)) - P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell | X \in R_{t,r}(\gamma;k))\Big)^2.$$

Here we still consider two cases: Because $k \notin U(t)$, either $(k,-1) \in \mathrm{F}^{\pm}(t)$ or $S_j^+ \cap \mathrm{F}^{\pm}(t) \neq \emptyset$.

i) If $S_j^+ \cap \mathrm{F}^{\pm}(t) \neq \emptyset$, suppose $(k',+1)$ is the first positive signed feature in S_j^+ that enters $\mathrm{F}^{\pm}(t)$. That means we can find a parent of t, denoted as t_{parent}, that splits on feature k' and none of t_{parent}'s parent splits on k'. That implies $k' \notin \mathrm{F}(t_{parent})$ and $S_j^+ \cap \mathrm{F}^{\pm}(t_{parent}) = \emptyset$, in other words, $k' \in U(t_{parent})$. Recall that $\gamma^*_{t_{parent},k'}$ denotes the threshold at node t_{parent}. By Lemma B.2.8, we know that the threshold $\gamma^*_{t_{parent},k'} \xrightarrow{p} \gamma_{k'}$. Since t is on the right branch of the node t_{parent}, $c_{low,k'}(t) \ge \gamma^*_{t_{parent},k'}$. Thus, $\mu(\{X|\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell\} \cap R_t) \xrightarrow{p} 0$. Since (B.68) is bounded by $2C_\beta^2 \frac{\mu(R_{t,l}(\gamma;k))\mu(R_{t,r}(\gamma;k))}{\mu(R_t)} \Big[P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell | X \in R_{t,l}(\gamma;k)) + P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell | X \in R_{t,r}(\gamma;k))\Big] \le 2C_\beta^2 P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_\ell, X \in R_t)$, we know that (B.68) converges to zero in probability.

ii) If $S_j^+ \cap \mathrm{F}^{\pm}(t) = \emptyset$ but $(k,-1) \in \mathrm{F}^{\pm}(t)$, it means there exists a parent of t, denoted t_{parent}, such that feature k is used to split that node and none of its parents splits on k, in other words, $k \in U(t_{parent})$. By Lemma B.2.8, we know that the corresponding threshold $\gamma^*_{t_{parent},k} \xrightarrow{p} \gamma_k$. Since $S_j^+ \cap \mathrm{F}^{\pm}(t) = \emptyset$, t is on the left branch of t_{parent}. Thus,

$c_{high,k}(t) \le \gamma^*_{t_{parent},k}$. For any fixed $\epsilon > 0$, if $\mu(R_{t,l}(\gamma;k)) > \epsilon$ and $\mu(R_{t,r}(\gamma;k)) > \epsilon$, then $P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_{\ell,j} | X \in R_{t,l}(\gamma;k)) - P(\forall\ \ell \in S_j,\ X_\ell \le \gamma_{\ell,j} | X \in R_{t,r}(\gamma;k)) \xrightarrow{p} 0$, which implies $\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \xrightarrow{p} 0$. Otherwise, $\mu(R_{t,l}(\gamma;k)) \le \epsilon$ or $\mu(R_{t,r}(\gamma;k)) \le \epsilon$, then $\frac{\mu(R_{t,l}(\gamma;k))\mu(R_{t,r}(\gamma;k))}{\mu(R_t)} \le \epsilon$ and $\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \le 4\epsilon$. Since ϵ is chosen arbitrarily, this implies $\Delta_I(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \xrightarrow{p} 0$.

Combining a) and b), we complete the proof.

□

With the help of the previous lemmas, we have the following proposition:

Proposition B.2.4. *Suppose t_{leaf} is a leaf of $\mathcal{P}$ from a random tree $T \in \mathcal{T}_2$. Assume A1-A5 hold true. For any fixed constant $\epsilon > 0$, the following facts hold true:*

i)

$$P\left(\max_{t\in T}\max_{k\in[p]/U(t)} \Delta_I^n(R_{t,l}(\gamma^*_{t,k};k), R_{t,r}(\gamma^*_{t,k};k)) < \frac{\epsilon}{4}C_\beta^2 C_\gamma^{2\max_j s_j - 1}\right) \to 1.$$

ii)

$$P\left(U(t_{leaf}) = \emptyset \middle| \mathcal{D}\right) \xrightarrow{p} 1.$$

iii)

$$P\left(\min_{t\in\mathfrak{p}(t_{leaf})}\min_{k\in U(t)} \Delta_I^n(R_{t,l}(\gamma^*_{t,k};k), R_{t,r}(\gamma^*_{t,k};k)) \ge \frac{\epsilon}{4}C_\beta^2 C_\gamma^{2\max_j s_j - 1} \middle| \mathcal{D}\right) \ge 1 - \epsilon^{\tilde{C}} - \eta_n(\mathcal{D},\epsilon),$$

with constant $\tilde{C} = C_m^{2s}/\log(1/C_\gamma)$ and $\eta_n(\mathcal{D},\epsilon) \xrightarrow{p} 0$.

Proof. i) By Lemma B.2.9, we know with probability approaching 1,

$$\max_{t\in T}\max_{k\in[p]/U(t)}\sup_{\gamma\in[0,1]} \Delta_I^n(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k)) \le \frac{\epsilon}{4}C_\beta^2 C_\gamma^{2\max_j s_j - 1}. \tag{B.69}$$

ii): For any fixed $\epsilon > 0$, by Lemma B.2.6 and Lemma B.2.9, the following event A_ϵ happens with probability approaching 1,

$$A_\epsilon = \bigcap_{T\in\mathcal{T}_1}\left\{\min_{t\in T,\mu(R_t)\ge\epsilon,U(t)\neq\emptyset}\min_{k\in U(t)}\sup_{\gamma\in[C_\gamma,1-C_\gamma]} \Delta_I^n(R_{t,l}(\gamma;k)\}, R_{t,r}(\gamma;k)) > \right. \tag{B.70}$$
$$\left.\max_{t\in T}\max_{k\in[p]/U(t)}\sup_{\gamma\in[0,1]} \Delta_I^n(R_{t,l}(\gamma;k), R_{t,r}(\gamma;k))\right\},$$

which implies that for any node with volume at least ϵ any desirable features has higher impurity decrease than any non-desirable feature. For a random path $\mathcal{P}$, denote its leaf node t_{leaf} and the depth of the path is D. Then for $d \in [D]$, denote t_d to be the d-th node on the

path $\mathcal{P}(t_{\text{leaf}})$. Recall that $S = \cup_{j=1}^{J} S_j$ denotes the set of all signal features and $s = |S|$ their total number. Based on (B.70), if at any node t, its candidate feature set $M_{\text{try}}(t)$ contains all the signal features S, then it will split on a signal feature as long as $U(t_{\text{leaf}}) \neq \emptyset$. If there are more than s nodes along the path that has volume larger than ϵ and their candidate feature set contains S, then the desirable features must have been exhausted at the leaf node, i.e.,

$$\left\{ \left| \{d \in [D] : S \subset M_{\text{try}}(t_d) \text{ and } \mu(R_d) \geq \epsilon\} \right| \geq s, A_\epsilon \right\} \subset \{U(t_{\text{leaf}}) = \emptyset, A_\epsilon\}. \tag{B.71}$$

Further, note that, because $\mu(R_{t_d}) \geq C_\gamma \mu(R_{t_{d-1}}) \geq \ldots \geq C_\gamma^d$, when $d < \log \epsilon / \log C_\gamma$, it always holds that $\mu(R_{t_{d-1}}) \geq \epsilon$ and therefore

$$\left\{ \left| \{d \in [\log \epsilon / \log C_\gamma] : S \subset M_{\text{try}}(t_d)\} \right| \geq s, A_\epsilon, D \geq \log \epsilon / \log C_\gamma \right\} \tag{B.72}$$

$$\subset \{U(t_{\text{leaf}}) = \emptyset, A_\epsilon, D \geq \log \epsilon / \log C_\gamma\}. \tag{B.73}$$

Since for any node t, its candidate feature set $M_{\text{try}}(t)$ has m_{try} features, we know

$$P(S \subset M_{\text{try}}(t)) = \frac{\binom{p-s}{m_{\text{try}}-s}}{\binom{p}{m_{\text{try}}}} = \frac{m_{\text{try}} \cdot (m_{\text{try}} - 1) \cdots (m_{\text{try}} - s + 1)}{p \cdot (p-1) \cdots (p-s+1)} \geq \left(\frac{m_{\text{try}} - s + 1}{p - s + 1} \right)^s \geq C_m^s. \tag{B.74}$$

Since $M_{\text{try}}(t)$ is independent of the path $\mathcal{P}$, it follows that

$$P_{(\mathcal{P},T)} \left(\left| \{d \in [\log \epsilon / \log C_\gamma] : S \subset M_{\text{try}}(t_d)\} \right| \geq s \middle| D \geq \log \epsilon / \log C_\gamma, \mathcal{D} \right) \tag{B.75}$$

$$\geq P(B(\log \epsilon / \log C_\gamma, C_m^s) \geq s) - \mathbf{1}(\mathcal{D} \in A_\epsilon) \tag{B.76}$$

$$\geq 1 - \exp\left(-2 \log \epsilon / \log C_\gamma \left(C_m^s - \frac{s}{\log \epsilon / \log C_\gamma} \right)^2 \right) - \mathbf{1}(\mathcal{D} \in A_\epsilon) \tag{B.77}$$

where $B(n,p)$ denotes a Binomial distribution with n trails and success probability p and the last inequaility follows from Hoeffding's inequality. Thus, for any $0 < \epsilon < \exp((1 - 1/\sqrt{2}) C_m^s / (s \log(1/C_\gamma)))$, we have

$$\left(C_m^s - \frac{s}{\log \epsilon / \log C_\gamma} \right)^2 \geq \frac{1}{2} C_m^{2s}.$$

Denote

$$\tilde{C} = C_m^{2s} / \log(1/C_\gamma),$$

we have that for sufficiently large n

$$P_{(\mathcal{P},T)} \left(\left| \{d \in [\log \epsilon / \log C_\gamma] : S \subset M_{\text{try}}(t_d)\} \right| \geq s \middle| D(\mathcal{P}) \geq \log \epsilon / \log C_\gamma, \mathcal{D} \right) \geq 1 - \epsilon^{\tilde{C}} - \mathbf{1}(\mathcal{D} \in A_\epsilon) \tag{B.78}$$

and thus it follows from (B.72) that

$$P_{(\mathcal{P},T)}\left(U(t_{\text{leaf}})=\emptyset\Big|D\geq \log\epsilon/\log C_\gamma,\mathcal{D}\right)\geq 1-\epsilon^{\tilde{C}}-\mathbf{1}(\mathcal{D}\in A_\epsilon). \tag{B.79}$$

Because $P(D\geq \log\epsilon/\log C_\gamma)\to 1$, by the Markov inequality, we know the random variable $P(D\geq \log\epsilon/\log C_\gamma\Big|\mathcal{D})\xrightarrow{p}1$. Thus, we know

$$P_{(\mathcal{P},T)}\left(U(t_{\text{leaf}})=\emptyset\Big|\mathcal{D}\right)\geq 1-\epsilon^{\tilde{C}}+\eta_n(\mathcal{D},\epsilon), \tag{B.80}$$

where $\eta_n(\mathcal{D},\epsilon)$ is a random variable only depend on $\mathcal{D}$ and $\eta_n(\mathcal{D},\epsilon)\xrightarrow{p}0$. Because that holds for any ϵ, we have

$$P_{(\mathcal{P},T)}\left(U(t_{\text{leaf}})=\emptyset\Big|\mathcal{D}\right)\xrightarrow{p}1.$$

iii) Denote t_s to be the s-th node in a path $\mathcal{P}(t_{\text{leaf}})$ for $s\geq 1$. Based on the proof of ii), let d be an integer that (roughly) equals to $\frac{\log\epsilon}{\log C_\gamma}$. Then $\mu(R_{t_d})\geq\epsilon$ and $P(U(t_d)=\emptyset\Big|\mathcal{D})\geq 1-\epsilon^{\tilde{C}}+\eta_n(\mathcal{D},\epsilon)$. When $U(t_d)=\emptyset$, $U(t_s)\neq\emptyset$ implies $s\leq d$ and $\mu(R_{t_s})\geq\epsilon$. Thus, $P(\exists t\in\mathcal{P}(t_{\text{leaf}}),\text{ such that } U(t)\neq\emptyset \text{ and } \mu(R_t)<\epsilon|\mathcal{D})\leq P(U(t_d)\neq\emptyset|\mathcal{D})\leq\epsilon^{\tilde{C}}-\eta_n(\mathcal{D},\epsilon)$. Therefore, we have

$$\begin{aligned}
&P\left(\min_{t\in\mathfrak{p}(t_{\text{leaf}}),U(t)\neq\emptyset}\min_{k\in U(t)}\Delta_I^n(R_{t,l}(\gamma_{t,k}^*;k),R_{t,r}(\gamma_{t,k}^*;k))\geq\frac{\epsilon}{4}C_\beta^2C_\gamma^{2\max_j s_j-1}\Big|\mathcal{D}\right)\\
\geq&P\left(\min_{t\in\mathfrak{p}(t_{\text{leaf}}),\mu(R_t)\geq\epsilon,U(t)\neq\emptyset}\min_{k\in U(t)}\Delta_I^n(R_{t,l}(\gamma_{t,k}^*;k),R_{t,r}(\gamma_{t,k}^*;k))\geq\frac{\epsilon}{4}C_\beta^2C_\gamma^{2\max_j s_j-1}\Big|\mathcal{D}\right)\\
&-\epsilon^{\tilde{C}}-\eta_n(\mathcal{D},\epsilon),
\end{aligned}$$

thus, the proof follows from Lemma B.2.8. □

Balanced root feature selection

Recall the definition of C_{root} in (B.26), which appears in Theorem B.2.1. Recall that for any tree T from RF, there are two different sources of randomness: the randomness of the data $\mathcal{D}=((\mathbf{x}_i,y_i))_{i=1}^n$ and the randomness from the candidate feature selection. Denote $M_{\text{try}}(t)\subset[p]$ to be the set of candidate features selected at node t and note that $M_{\text{try}}(t)$ and the data $\mathcal{D}$ are independent.

Define the event A to be that, given data $\mathcal{D}$, the maximum impurity decrease at the split of root node for every signal feature $k\in\cup_j S_j$ is larger than that of any noisy feature $k'\notin\cup_j S_j$, that is,

$$A=\{\min_{k\in\cup_j S_j}\Delta_I^n\left(R_{t_{\text{root}},l}(\gamma_k^\star,k),R_{t_{\text{root}},r}(\gamma_k^\star,k)\right)>\max_{k'\notin\cup_j S_j}\Delta_I^n\left(R_{t_{\text{root}},l}(\gamma_{k'}^\star,k'),R_{t_{\text{root}},r}(\gamma_{k'}^\star,k')\right)\}\in\sigma(\mathcal{D}). \tag{B.81}$$

Here $\sigma(\mathcal{D})$ is the sigma field induced by $\mathcal{D}$. By definition A is independent of $M_{\text{try}}(t_{\text{root}})$. Note that it follows from Proposition B.2.4 that

$$P_{\mathcal{D}}(A)\to 1\quad\text{ as } n\to\infty.$$

Theorem B.2.2. *Assume that $C_m p \leq m_{try} \leq (1 - C_m)(p - s + 1) + 1$ for some constant $C_m \in (0, 1)$. Condition on $\mathcal{D} = ((\mathbf{x}_i, y_i))_{i=1}^n$, for any $k \in \cup_j S_j$, we have*

$$P_T\left(t_{\text{root}} \text{ splits on feature } k|\mathcal{D}\right) \geq C_m^s - 1_{A^c}$$

and thus,

$$C_{\text{root}}(\mathcal{D}) \geq C_m^s - 1_{A^c} \xrightarrow{p} C_m^s \quad \text{as } n \to \infty.$$

Proof. For any $k \in \cup_j S_j$, define B_k to be the event that only signal feature k is selected in $M_{\text{try}}(t_{\text{root}})$

$$B_k \triangleq \{M_{\text{try}}(t_{\text{root}}) \cap \cup_j S_j = k \text{ and } |M_{\text{try}}(t_{\text{root}}) \setminus \cup_j S_j| = m_{\text{try}} - 1\}.$$

B_k only depends on $M_{\text{try}}(t_{\text{root}})$ but does not depend on $\mathcal{D}$. Note that

$$A \cap B_k \subset \{t_{\text{root}} \text{ splits on feature } k\}.$$

Thus,

$$P_T\left(t_{\text{root}} \text{ splits on feature } k|\mathcal{D}\right) \geq P_T(B_k \cap A|\mathcal{D}) \geq P_T\left(B_k|\mathcal{D}\right) - P_T\left(A^c|\mathcal{D}\right) = P\left(B_k\right) - 1_{A^c}.$$

Moreover, we have that

$$\begin{aligned} P(B_k) &= \frac{\binom{p-s}{m_{\text{try}}-1}}{\binom{p}{m_{\text{try}}}} = \frac{m_{\text{try}}}{p}\frac{\binom{p-s}{m_{\text{try}}-1}}{\binom{p-1}{m_{\text{try}}-1}} = \frac{m_{\text{try}}}{p}\frac{\binom{p-m_{\text{try}}}{s-1}}{\binom{p-1}{s-1}} \\ &= \prod_{i=0}^{s-2}\left(\frac{p - m_{\text{try}} - i}{p - 1 - i}\right)\frac{m_{\text{try}}}{p} \geq \left(\frac{p - m_{\text{try}} - s + 2}{p - s + 1}\right)^{s-1}\frac{m_{\text{try}}}{p} \geq C_m^s \end{aligned}$$

where the second equality follows from the identity where $n = p - 1$, $h = s - 1$, and $k = m_{\text{try}} - 1$:

$$\frac{\binom{n-h}{k}}{\binom{n}{k}} = \frac{\binom{n-k}{h}}{\binom{n}{h}}.$$

□

Note that Theorem B.2.2 suggests that we want to chose m_{try} such that C_m is as large as possible. From the constraint $C_m p \leq m_{\text{try}} \leq (1 - C_m)(p - s + 1) + 1$ we obtain that $C_m \leq (p - s + 2)/(2p - s + 1)$, where equality corresponds to the choice

$$m_{\text{try}} = \frac{1 - \frac{s-2}{p}}{2 - \frac{s-1}{p}}\, p. \tag{B.82}$$

When $p \gg s$, $m_{\text{try}} \approx p/2$ and $C_m \approx 1/2$.

Combining results

Our major result in Theorem B.2.1 is formulated for the random (oracle) feature set $\mathcal{F} = \mathcal{F}(\mathcal{D}, T, \mathcal{P})$. Note that this is an oracle feature set, as it depends on the true interactions S_j, which are not known in practice. Further, note that given a tree $T = T(\mathcal{D})$, we have that $\mathcal{F}$ is independent of the data $\mathcal{D}$, that is

$$\mathcal{F} \mid T \perp\!\!\!\perp \mathcal{D}. \tag{B.83}$$

From the analysis in Section B.2 we know that we can obtain a consistent estimate of the oracle feature set $\mathcal{F}$ by thresholding on MDI as in $\hat{\mathcal{F}}_\epsilon$. Recall that for a given ϵ the (random) set $\hat{\mathcal{F}}_\epsilon$ can easily be obtained without any knowledge of the true model. Also, note that the property (B.83) does not hold for $\hat{\mathcal{F}}_\epsilon$. Based on Proposition B.2.4, we observe the following.

Recall that Ω_0 is defined in (B.24), $\mathcal{F}$ is defined in (B.23), and $\hat{\mathcal{F}}_\epsilon$ is defined in (9.5) in the main text. We have the following theorem.

Theorem B.2.3. *Under the assumption of Proposition B.2.4 it holds true that for any fixed* $\epsilon > 0$,

$$P_{(\mathcal{P},T)}\left(\Omega_0^c \mid \mathcal{D}\right) \xrightarrow{p} 0; \tag{B.84}$$

$$P_{(\mathcal{P},T)}\left(\hat{\mathcal{F}}_\epsilon \nsubseteq \mathcal{F} \;\middle|\; \mathcal{D}\right) \xrightarrow{p} 0 \tag{B.85}$$

$$P_{(\mathcal{P},T)}\left(\hat{\mathcal{F}}_\epsilon \neq \mathcal{F} \;\middle|\; \mathcal{D}\right) \leq \left(\frac{4\epsilon}{C_\beta^2 C_\gamma^{2\max_j s_j - 1}}\right)^{\tilde{C}} + \eta_n(\mathcal{D}, \epsilon) \quad \text{with } \eta_n(\mathcal{D}, \epsilon) \xrightarrow{p} 0; \tag{B.86}$$

with $\tilde{C}$ *as in Proposition B.2.4.*

Proof. (B.84) follows directly from Proposition B.2.4 ii) and the definition of Ω_0 in (B.24).

To prove (B.85), one observes from Proposition B.2.4 i) that for any $\epsilon > 0$, taking $\tilde{\epsilon} = \frac{4\epsilon}{C_\beta^2 C_\gamma^{2\max_j s_j - 1}}$, the following happens with probability converging to one (as $n \to \infty$)

$$\max_{t \in T} \max_{k \in [p]/U(t)} \Delta_I^n(R_{t,l}(\gamma_{t,k}^*; k), R_{t,r}(\gamma_{t,k}^*; k)) < \frac{\tilde{\epsilon}}{4} C_\beta^2 C_\gamma^{2\max_j s_j - 1} = \epsilon,$$

which implies that $\hat{\mathcal{F}}_\epsilon$ contains no irrelevant features. Thus,

$$\liminf_{n \to \infty} P_{(\mathcal{D},T,\mathcal{P})}\left(\hat{\mathcal{F}}_\epsilon \subseteq \mathcal{F}\right) = 1.$$

Then by Markov inequality, we know $P_{(\mathcal{P},T)}\left(\hat{\mathcal{F}}_\epsilon \nsubseteq \mathcal{F} \;\middle|\; \mathcal{D}\right) \xrightarrow{p} 0$.

To prove (B.86), we further note that by Proposition B.2.4 iii),

$$P\left(\min_{t \in \mathcal{P}(t_{\text{leaf}})} \min_{k \in U(t)} \Delta_I^n(R_{t,l}(\gamma_{t,k}^*; k), R_{t,r}(\gamma_{t,k}^*; k)) \geq \epsilon \middle| \mathcal{D}\right) \geq 1 - \left(\frac{4\epsilon}{C_\beta^2 C_\gamma^{2\max_j s_j - 1}}\right)^{\tilde{C}} - \eta_n(\mathcal{D}, \epsilon).$$

If

$$\min_{t\in\mathfrak{p}(t_{\text{leaf}})}\min_{k\in U(t)}\Delta_I^n(R_{t,l}(\gamma^*_{t,k};k),R_{t,r}(\gamma^*_{t,k};k))\geq\epsilon$$

and

$$\max_{t\in T}\max_{k\in[p]/U(t)}\Delta_I^n(R_{t,l}(\gamma^*_{t,k};k),R_{t,r}(\gamma^*_{t,k};k))<\epsilon$$

, we know $\hat{\mathcal{F}}_\epsilon=\mathcal{F}$. Thus, we have

$$P_{(T,\mathcal{P})}\left(\hat{\mathcal{F}}_\epsilon=\mathcal{F}\middle|\mathcal{D}\right)\geq 1-\left(\frac{4\epsilon}{C_\beta^2C_\gamma^{2\max_j s_j-1}}\right)^{\tilde{C}}-\eta_n(\mathcal{D},\epsilon). \tag{B.87}$$

That completes the proof. □

Finally, we can combine Theorem B.2.1, Theorem B.2.2, and Theorem B.2.3 to prove Theorem 9.4.1 in the main text.

Proof of Theorem 9.4.1. Assume that $|S^\pm|=\tilde{s}$ and $S^\pm=\{(k_1,b_1),\ldots,(k_{\tilde{s}},b_{\tilde{s}})\})\}$ and let

$$r_n(\mathcal{D},\epsilon)=\max\left(P_{(\mathcal{P},T)}(\Omega_0^c|\mathcal{D})+\eta_n(\mathcal{D},\epsilon),P_{(\mathcal{P},T)}(\hat{\mathcal{F}}_\epsilon\not\subseteq\mathcal{F}|\mathcal{D})\right),$$

with $\eta_n(\mathcal{D},\epsilon)$ as in Theorem B.2.3. It follows from Theorem B.2.3 that $r_n(\mathcal{D},\epsilon)\xrightarrow{p}0$ as $n\to\infty$.

Proof of 1.:

Analog as in Theorem B.2.1, for any feature $k\in[p]$, let B^k be the Bernoulli random variable we draw when k appears for the first time on $\mathcal{P}$. Recall the definition of $\hat{\mathcal{F}}_\epsilon$, in particular, that $(k,b_k)\in\hat{\mathcal{F}}_\epsilon$ only if X_k appears the first time on $\mathcal{P}$. Thus, analog as for $\mathcal{F}$ (recall the proof of Theorem B.2.1) we have that $(k,-1)\in\mathcal{F}$ implies $B^k=-1$ and $(k,+1)\in\mathcal{F}$ implies $B^k=+1$. Thus,

$$\{S^\pm\in\hat{\mathcal{F}}_\epsilon\}\subset\{B^{k_1}=b_1\cap\ldots\cap B^{k_{\tilde{s}}}=b_{\tilde{s}}\}$$

and hence,

$$\begin{aligned}\text{DWP}(S^\pm)=&P_{(\mathcal{P},T)}(S^\pm\in\hat{\mathcal{F}}_\epsilon|\mathcal{D})\\ \leq&P_{(\mathcal{P},T)}(B^{k_1}=b_1\cap\ldots\cap B^{k_{\tilde{s}}}=b_{\tilde{s}}|\mathcal{D})\\ =&P_{\mathcal{P}}(B^{k_1}=b_1\cap\ldots B^{k_{\tilde{s}}}=b_{\tilde{s}})=2^{-\tilde{s}}.\end{aligned}$$

Proof of 2.:
Assume that $S^{\pm}$ is a union interaction. Then we have that

$$\begin{aligned}
\mathrm{DWP}(S^{\pm}) &= P_{(\mathcal{P},T)}(S^{\pm} \in \hat{\mathcal{F}}_{\epsilon}|\mathcal{D}) \\
&\geq P_{(\mathcal{P},T)}(S^{\pm} \in \mathcal{F}|\mathcal{D}) - P_{(\mathcal{P},T)}(\hat{\mathcal{F}}_{\epsilon} \neq \mathcal{F}|\mathcal{D}) \\
&\geq P_{(\mathcal{P},T)}(S^{\pm} \in \mathcal{F}|\mathcal{D}) - \left(\frac{4\epsilon}{C_{\beta}^{2} C_{\gamma}^{2\max_j s_j - 1}}\right)^{\tilde{C}} - \eta_n(\mathcal{D}, \epsilon) \\
&\geq 0.5^{\tilde{s}} - P_{(\mathcal{P},T)}(\Omega_0^c|\mathcal{D}) - \left(\frac{4\epsilon}{C_{\beta}^{2} C_{\gamma}^{2\max_j s_j - 1}}\right)^{\tilde{C}} - \eta_n(\mathcal{D}, \epsilon) \\
&\geq 0.5^{\tilde{s}} - \left(\frac{4\epsilon}{C_{\beta}^{2} C_{\gamma}^{2\max_j s_j - 1}}\right)^{\tilde{C}} - r_n(\mathcal{D}, \epsilon),
\end{aligned}$$

where the second inequality follows from Corollary B.2.3 and the third inequality follows from Theorem B.2.1.
Proof of 3.:
Assume that $S^{\pm}$ is not a union interaction. Then we have that

$$\begin{aligned}
\mathrm{DWP}(S^{\pm}) &= P_{(\mathcal{P},T)}(S^{\pm} \in \hat{\mathcal{F}}_{\epsilon}|\mathcal{D}) \\
&\leq P_{(\mathcal{P},T)}(S^{\pm} \in \mathcal{F}|\mathcal{D}) + P_{(\mathcal{P},T)}(\hat{\mathcal{F}}_{\epsilon} \nsubseteq \mathcal{F}|\mathcal{D}) \\
&\leq 0.5^{\tilde{s}}(1 - C_{\mathrm{root}}(\mathcal{D})/2) + r_n(\mathcal{D}, \epsilon),
\end{aligned}$$

where the second inequality follows from Theorem B.2.2. □

Bibliography

[1] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. "A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries". In: *arXiv preprint* arxiv:1309.1952 (2013), pp. 1–31. ISSN: 0018-9448. arXiv: 1309.1952.

[2] Alekh Agarwal et al. "Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization". In: *SIAM Journal on Optimization* 26.4 (2014), pp. 2775–2799.

[3] Michal Aharon, Michael Elad, and Alfred Bruckstein. "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation". In: *IEEE Transactions on Signal Processing* 54.11 (2006), pp. 4311–4322. ISSN: 1053587X. arXiv: 59749104367.

[4] Michal Aharon, Michael Elad, and Alfred M Bruckstein. "K-SVD and its Non-Negative Variant for Dictionary Design". In: *International Society for Optics and Photonics* 5914 (2005), p. 591411.

[5] Sanjeev Arora, Rong Ge, and Ankur Moitra. "New algorithms for learning incoherent and overcomplete dictionaries". In: *Conference on Learning Theory*. 2014, pp. 779–806.

[6] Sanjeev Arora et al. "More algorithms for provable dictionary learning". In: *arXiv preprint arXiv:1401.0579* (2014), p. 23. arXiv: 1401.0579.

[7] Sanjeev Arora et al. "Simple, Efficient, and Neural Algorithms for Sparse Coding". In: *arXiv:1503.00778 [cs, stat]* (2015). ISSN: 15337928. arXiv: 1503.00778.

[8] A. M. Bagirov et al. "Subgradient Method for Nonconvex Nonsmooth Optimization". In: *Journal of Optimization Theory and Applications* 157.2 (2013), pp. 416–435. ISSN: 1573-2878. DOI: 10.1007/s10957-012-0167-6. URL: https://doi.org/10.1007/s10957-012-0167-6.

[9] Boaz Barak, Jonathan A Kelner, and David Steurer. "Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method". In: *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*. STOC '15 (2014), pp. 143–151. ISSN: 07378017. arXiv: 1407.1543.

[10] Sumanta Basu et al. "Iterative random forests to discover predictive and stable high-order interactions". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.8 (2018), pp. 1943–1948. ISSN: 10916490. eprint: 1706.08457.

[11] Gérard Biau. "Analysis of a random forests model". In: *Journal of Machine Learning Research* 1 (2012). ISSN: 15324435.

[12] Leo Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 1–33.

[13] L Breiman et al. *Classification and Regression Trees*. New York: Chapman and Hall, 1984.

[14] Jean-Philippe Brunet et al. "Metagenes and molecular pattern discovery using matrix factorization". In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169. ISSN: 0027-8424.

[15] Emmanuel J. Candes, Michael B. Wakin, and Stephen P. Boyd. "Enhancing sparsity by reweighted L1 minimization". In: *Journal of Fourier analysis and applications* 14.5 (2008), pp. 877–905.

[16] Strobl Carolin, Hothorn Torsten, and Zeileis Achim. "Party on! A New, Conditional Variable-Importance Measure for Random Forests Available in the party Package". In: *the R journal* 1/2 (2009), pp. 14–17.

[17] Susan E Celniker et al. "Unlocking the secrets of the genome". In: *Nature* 459.7249 (2009), p. 927.

[18] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. "Atomic decomposition by basis pursuit". In: *SIAM review* 43.1 (2001), pp. 129–159.

[19] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. arXiv: 1603.02754.

[20] Xi Chen and Hemant Ishwaran. "Random forests for genomic data analysis". In: *Genomics* 99.6 (2012), pp. 323–329.

[21] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[22] Misha Denil, David Matheson, and Nando De Freitas. "Narrowing the Gap: Random Forests In Theory and In Practice". In: ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 1. Bejing, China: PMLR, June 2014, pp. 665–673.

[23] Anulekha Dhara and Joydeep Dutta. *Optimality conditions in convex optimization: a finite-dimensional view*. CRC Press, 2011.

[24] R Diaz-Uriarte and S de Andrés. "Gene Selection and Classification of Microarray Data Using Random Forest". In: *BMC Bioinformatics* 7 (2006).

[25] Michael Elad and Michal Aharon. "Image denoising via sparse and redundant representations over learned dictionaries". In: *Image Processing, IEEE Transactions on* 15.12 (2006), pp. 3736–3745.

[26] James E. Ferrell Jr. "Tripping the Switch Fantastic: How a Protein Kinase Cascade Can Convert Graded Inputs into Switch-like Outputs". In: *Trends in Biochemical Sciences* 21.12 (1996), pp. 460–466.

[27] Jerome H Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[28] J-J Fuchs. "On sparse representations in arbitrary redundant bases". In: *IEEE transactions on Information theory* 50.6 (2004), pp. 1341–1344.

[29] Rong Ge, Jason D. Lee, and Tengyu Ma. "Matrix Completion has No Spurious Local Minimum". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems* (2016), pp. 1–27. ISSN: 10495258. arXiv: 1605.07272.

[30] Quan Geng, John Wright, and Huan Wang. "On the local correctness of L1-minimization for dictionary learning". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 3180–3184.

[31] Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. "Sparse and Spurious: Dictionary Learning With Noise and Outliers". In: *IEEE Transactions on Information Theory* (2015). ISSN: 00189448. arXiv: 1407.5155.

[32] Rémi Gribonval and Karin Schnass. "Dictionary Identification - Sparse Matrix-Factorisation via L1-Minimisation". In: *Information Theory, IEEE Transactions on* 56.7 (2010), pp. 3523–3539.

[33] Ann S Hammonds et al. "Spatial expression of transcription factors in Drosophila embryonic organ development". In: *Genome biology* 14.12 (2013), R140.

[34] Sepp Hochreiter and Jürgen Schmidhuber. "Flat minima". In: *Neural Computation* 9.1 (1997), pp. 1–42.

[35] Michael M Hoffman et al. "Integrative annotation of chromatin elements from ENCODE data". In: *Nucleic acids research* 41.2 (2013), pp. 827–841.

[36] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15 (2006).

[37] Patrik O. Hoyer. "Non-negative sparse coding". In: *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop* 2002-Janua (2002), pp. 557–565. ISSN: 0780376161. arXiv: 0202009 [cs].

[38] Vân Anh Huynh-Thu et al. "Inferring regulatory networks from expression data using tree-based methods". In: *PLoS ONE* 5.9 (2010). ISSN: 19326203.

[39] Silke Janitza, Ender Celik, and Anne Laure Boulesteix. "A computationally fast variable importance test for random forests for high-dimensional data". In: *Advances in Data Analysis and Classification* 12.4 (2016), pp. 1–31. ISSN: 18625355.

[40] Peng Jiang et al. "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features". In: *Nucleic acids research* 35.suppl_2 (2007), W339–W344.

[41] Jalil Kazemitabar et al. "Variable importance using decision trees". In: *Advances in Neural Information Processing Systems*. 2017, pp. 426–435.

[42] Oren Kobiler et al. "Quantitative Kinetic Analysis of the Bacteriophage Genetic Network". In: *Proceedings of the National Academy of Sciences* 102.12 (Mar. 2005), pp. 4470–4475.

[43] Kenneth Kreutz-delgado et al. "Dictionary learning algorithms for sparse representation." In: *Neural computation* 15.2 (2003), pp. 349–96. ISSN: 0899-7667. arXiv: /www.ncbi.nlm.nih.gov/pmc/articles/PMC2944020/pdf/nihms234072.pdf [http:].

[44] Karl Kumbier et al. "Refining interaction search through signed iterative Random Forests". In: *arXiv preprint arXiv:1810.07287* (2018).

[45] Karl Kumbier et al. "Refining interaction search through signed iterative Random Forests". In: *bioRxiv* 1 (2018). eprint: https://www.biorxiv.org/content/early/2018/11/11/467498.full.pdf.

[46] Daniel Lee and Sebastian Seung. "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems 13*. 2001. arXiv: 0408058v1 [arXiv:cs].

[47] Jung Bok Jae Won Lee et al. "An extensive comparison of recent classification tools applied to microarray data". In: *Computational Statistics and Data Analysis* 48.4 (2005), pp. 869–885. ISSN: 01679473.

[48] Sylvain Lesage et al. "Learning unions of orthonormal bases with thresholded singular value decomposition". In: *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Vol. 5. IEEE. IEEE, 2005, pp. v–293.

[49] Erel Levine and Terence Hwa. "Small RNAs Establish Gene Expression Thresholds". In: *Current Opinion in Microbiology* 11.6 (Dec. 2008), pp. 574–579.

[50] Xiao Li et al. "A Debiased MDI Feature Importance Measure for Random Forests". In: *Advances in Neural Information Processing Systems*. San Diego, 2019, pp. 8047–57. arXiv: 1906.10845.

[51] J. W. Little. "Threshold Effects in Gene Regulation: When Some Is Not Enough". In: *Proceedings of the National Academy of Sciences* 102.15 (Apr. 2005), pp. 5310–5311.

[52] J. W. Little, Shepley, and Wert. "Robustness of a Gene Regulatory Circuit". In: *The EMBO Journal* 18.15 (Aug. 1999), pp. 4299–4307.

[53] Markus Loecher. "Unbiased Variable Importance for Random Forests". In: *Communications in Statistics - Theory and Methods* (May 2020), pp. 1–13. arXiv: 2003.02106.

[54] Wei-Yin Loh. "Fifty years of classification and regression trees". In: *International Statistical Review* 82.3 (2014), pp. 329–348. ISSN: 17515823.

[55] Gilles Louppe. "Understanding random forests: From theory to practice". In: *arXiv preprint arXiv:1407.7502* (2014).

[56] Gilles Louppe et al. "Understanding variable importances in forests of randomized trees". In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 431–439.

[57] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: *ArXiv e-prints arXiv:1802.03888* (2018). arXiv: 1802.03888.

[58] Stewart MacArthur et al. "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions". In: *Genome biology* 10.7 (2009), p. 1.

[59] Julien Mairal et al. "Online Dictionary Learning for Sparse Coding". In: *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), pp. 689–696.

[60] Julien Mairal et al. *SPAMS : a SPArse Modeling Software , v2.5*. 2014.

[61] Julien Mairal et al. "Supervised dictionary learning". In: *Advances in neural information processing systems*. 2009, pp. 1033–1040.

[62] W. James Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080. ISSN: 0027-8424. eprint: https://www.pnas.org/content/116/44/22071.full.pdf.

[63] Saralees Nadarajah and Samuel Kotz. "On The Linear Combination Of Laplace Random Variables". In: *Probab. Eng. Inf. Sci.* 19.4 (Oct. 2005), pp. 463–470. ISSN: 0269-9648.

[64] Stefano Nembrini, Inke R. König, and Marvin N. Wright. "The revival of the Gini importance?" In: *Bioinformatics* 34.21 (2018), pp. 3711–3718. ISSN: 14602059.

[65] Bruno A. Olshausen and David J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609.

[66] Bruno A. Olshausen and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* 37.23 (1997), pp. 3311–3325. ISSN: 00426989. arXiv: 1409.1556.

[67] Jason T. Parker and Philip Schniter. "Parametric Bilinear Generalized Approximate Message Passing". In: *IEEE Journal on Selected Topics in Signal Processing*. Vol. 10. 2016, pp. 795–808. arXiv: arXiv:1310.2632v2.

[68] Jason T. Parker, Philip Schniter, and Volkan Cevher. "Bilinear generalized approximate message passing—Part I: Derivation". In: *IEEE Transactions on Signal Processing* 62.22 (2014), pp. 5839–5853.

[69] Jason T. Parker, Philip Schniter, and Volkan Cevher. "Bilinear generalized approximate message passing—Part II: Applications". In: *IEEE Transactions on Signal Processing* 62.22 (2014), pp. 5854–5867.

[70] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[71] Gabriel Peyré. "Sparse modeling of textures". In: *Journal of Mathematical Imaging and Vision* 34.1 (2009), pp. 17–31.

[72] Boris Teodorovic Polyak. "Sharp Minima". In: *Institute of Control Sciences Lecture Notes, Moscow IIASA Workshop On Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria*. 1979.

[73] Qiang Qiu, Vishal M Patel, and Rama Chellappa. "Information-theoretic Dictionary Learning for Image Classification". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.11 (2014), pp. 2173–2184.

[74] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?" Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016).

[75] Wendy Rodenburg et al. "A Framework to Identify Physiological Responses in Microarray Based Gene Expression Studies: Selection and Interpretation of Biologically Relevant Genes". In: *Physiological Genomics* 33 (2008).

[76] Joseph Lee Rodgers, W. Alan Nicewander, and Larry Toothaker. "Linearly independent, orthogonal, and uncorrelated variables". In: *American Statistician* 38.2 (1984), pp. 133–134. ISSN: 15372731.

[77] Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. "Dictionaries for sparse representation modeling". In: *Proceedings of the IEEE* 98.6 (2010), pp. 1045–1057. ISSN: 00189219. arXiv: 0402594v3 [arXiv:cond-mat].

[78] Ando Saabas. *Interpreting random forests*. 2014.

[79] Marco Sandri and Paola Zuccolotto. "A bias correction algorithm for the gini variable importance measure in classification trees". In: *Journal of Computational and Graphical Statistics* 17.3 (2008), pp. 611–628. ISSN: 10618600.

[80] Karin Schnass. "Local Identification of Overcomplete Dictionaries". In: *Journal of Machine Learning Research* 16 (2015), pp. 1211–1242. ISSN: 15337928. arXiv: 1401.6354.

[81] Karin Schnass. "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD". In: *Applied and Computational Harmonic Analysis* 37.3 (2014), pp. 464–491. ISSN: 1096603X. arXiv: 1301.3375.

[82] Erwan Scornet. "Tuning parameters in random forests". In: *ESAIM: Proceedings and Surveys* 60 (2017), pp. 144–162.

[83] Erwan Scornet, Gerard Biau, and Jean Philippe Vert. "Consistency of random forests". In: *Annals of Statistics* 43.4 (2015), pp. 1716–1741. ISSN: 00905364. arXiv: 1405.2881.

[84] Daniel A. Spielman, Huan Wang, and John Wright. "Exact recovery of sparsely-used dictionaries". In: *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press. 2013, pp. 3087–3090. arXiv: 1206.5882.

[85] Nathan Srebro and Tommi Jaakkola. "Weighted low-rank approximations". In: *Proceedings of the Twentieth International Conference on Machine Learning*. Vol. 3. 2003, pp. 720–727.

[86] Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. "Unbiased split selection for classification trees based on the Gini index". In: *Computational Statistics & Data Analysis* 52.1 (2007), pp. 483–501.

[87] Carolin Strobl et al. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution". In: *BMC Bioinformatics* 8 (2007).

[88] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1 (2008), p. 307. ISSN: 1471-2105.

[89] Erik Strumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and Information Systems* 41.3 (2014), pp. 647–665. ISSN: 02193116.

[90] Ju Sun, Qing Qu, and John Wright. "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 853–884. ISSN: 00189448. arXiv: 1511.03607.

[91] Ju Sun, Qing Qu, and John Wright. "Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust-Region Method". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 885–914. ISSN: 00189448. arXiv: 1504.06785.

[92] Wouter G. Touw et al. "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?" In: *Briefings in Bioinformatics* 14.3 (July 2012), pp. 315–326. ISSN: 1467-5463. DOI: 10.1093/bib/bbs034.

[93] Vladimir Vapnik. *Statistical learning theory*. 1998.

[94] Stefan Wager and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests". In: *Journal of the American Statistical Association* 113.523 (July 2018), pp. 1228–1242.

[95] Xiang Wan et al. "MegaSNPHunter: A Learning Approach to Detect Disease Predisposition SNPs and High Level Interactions in Genome Wide Association Study". In: *BMC Bioinformatics* 10.1 (Jan. 2009), p. 13.

[96] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. "Variable importance analysis: A comprehensive review". In: *Reliability Engineering and System Safety* 142 (2015), pp. 399–432. ISSN: 09518320.

[97] Christoph Witzgall and R. Fletcher. "Practical Methods of Optimization." In: *Mathematics of Computation* (1989). ISSN: 00255718. arXiv: 1003.3921v1.

[98] Lewis Wolpert. "Positional information and the spatial pattern of cellular differentiation". In: *Journal of theoretical biology* 25.1 (1969), pp. 1–47.

[99] Marvin Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software, Articles* 77.1 (2017), pp. 1–17. ISSN: 1548-7660. URL: https://www.jstatsoft.org/v077/i01.

[100] Siqi Wu and Bin Yu. "Local identifiability of L1-minimization dictionary learning: a sufficient and almost necessary condition". In: *Journal of Machine Learning Research* 18 (2018), pp. 1–56.

[101] Siqi Wu et al. "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks". In: *Proceedings of the National Academy of Sciences* 113.16 (2016), p. 201521171. ISSN: 0027-8424.

[102] Makiko Yoshida and Asako Koike. "SNPInterForest: A New Method for Detecting Epistatic Interactions". In: *BMC Bioinformatics* 12.1 (Dec. 2011).

[103] Zhengze Zhou and Giles Hooker. "Unbiased Measurement of Feature Importance in Tree-Based Methods". In: *arXiv:1903.05179 [cs, stat]* (Mar. 2020). arXiv: 1903.05179 [cs, stat].

[104] Michael Zibulevsky and Barak A. Pearlmutter. "Blind Source Separation by Sparse Decomposition in a Signal Dictionary". In: *Neural computation* 4.13 (2001), pp. 863–882.