UNIVERSITY OF CALIFORNIA, SAN DIEGO

Duality and Data Dependence in Boosting

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

in

Computer Science

by

Matus Telgarsky

Committee in charge:

Professor Sanjoy Dasgupta, Chair
Professor Kamalika Chaudhuri
Professor Patrick Fitzsimmons
Professor Yoav Freund
Professor Philip Gill
Professor Robert Schapire

2013

The Dissertation of Matus Telgarsky is approved and is acceptable in quality and form for publication on micro lm and electronically:

Chair

University of California, San Diego

2013

EPIGRAPH

In mathematics you don't understand things. You just get used to them.

John von Neumann

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Chapters 1, 2, and 3 contain material from the Journal of Machine Learning Research, volume 13, pages 561{606, 2012. The dissertation author was the sole author of this paper.

Chapters 1, 4, and 6 contain material from the Conference on Learning Theory, 2013. The dissertation author was the sole author of this paper.

Chapters 1 and 5 contains material being prepared for submission. The dissertation author is the sole author.

Much of this work was supported by the NSF under grants IIS-0713540 and IIS-0812598.

The author thanks the following people for valuable discussions throughout the project: Jake Abernethy, Akshay Balsubramani, Sanjoy Dasgupta, Daniel Hsu, Brian McFee, Aditya Menon, Indraneel Mukherjee, Alexander Rakhlin, Jeroen Rombouts, Robert Schapire, Karthik Sridharan, Manfred Warmuth, and numerous reviewers. The author wishes to give special mention to: his advisor Sanjoy Dasgupta for support and mathematical nutrition; Robert Schapire for insight, support, discussions, and for sharing open problems, early book drafts, and enthusiasm; Daniel Hsu for insight, support, discussions, for sparking the author's interest in boosting by initiating a reading project, and for setting the research path of this thesis by suggesting the author solve the problem in Chapter 3; his mother, Anna Telgarsky, for a work ethic.

VITA

2003	Diploma in Violin Performance, The Juilliard School
2007	B.S. in Computer Science and Discrete Math, Carnegie Mellon University
2013	Ph.D. in Computer Science, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Duality and Data Dependence in Boosting

by

Matus Telgarsky

Doctor of Philosophy in Computer Science

University of California, San Diego, 2013

Professor Sanjoy Dasgupta, Chair

Boosting algorithms produce accurate predictors for complex phenomena by welding together collections of simple predictors. In the classical method AdaBoost, as well as its immediate variants, the welding points are determined by convex optimization; unlike typical applications of convex optimization in machine learning, however, the AdaBoost scheme eschews the usual regularization and constraints used to control numerical and statistical properties.

On the other hand, the data and simple predictors impose rigid structure on the behavior of AdaBoost variants, and moreover convex duality provides a lens to resolve this rigidity. This structure is fundamental to the properties of these methods, and in particular leads to numerical and statistical convergence rates.

Chapter 1

Introduction

1.1 Boosting

Let a room with heaps of books be given; since the niteness of human life precludes reading all of them, what is the best way to identify a high quality subset? It is unclear how to de ne a good book. On the other hand, it is easy to identify many characteristics which are correlated with the quality of a book: for example, it should have spaceships, and it should contain information, but reading it should not lead to sadness for no reason.

This is precisely the idea behind*boosting*: rather than carefully reasoning about an appropriate model for a prediction problem, simply identify a set of characteristics correlated with the intended values, and leave it to an algorithm to combine these scraps of information into an accurate model.

The key question here is of course what such a scheme can hope to achieve. An early concrete question along these lines was provided by Kearns and Valiant (1989), where the task was to identify individual books as good or bad (for an abstract notion of book). In more detail, suppose there exists a positive constant > 0 so that for any distribution over the set of books, some characteristic has correlation at least with the quality of the books according to the selected distribution; is it possible to construct a circuit which uses correlation values for a set of books to separate the good ones from the bad?

The de nition of suggests an iterative scheme, since an initial circuit gives rise to a distribution emphasizing those books where the circuit errs, which in turn by the stated assumption leads to a characteristic that is -correlated with book quality. If this characteristic can be baked into the initial circuit in such a way that accuracy improves, then perhaps the goal

may be achieved by recursing.

The landmark result here, provided by Schapire (1990), is that this can indeed be done. The circuit provided by Schapire's boosting method was a ternary tree, where leaves measured correlation values for a provided book, and internal nodes took a majority vote over their three children.

The next major result, by Freund (1995), was that not only did a single majority over a set of correlation values su ce, but moreover only $O(\ln(1/\varepsilon)/\gamma^2)$ correlates were used, where $\gamma > 0$ is the target accuracy (e.g., choosing $\gamma < 1/(\text{number of books})$ su ces to classify the books perfectly). This scheme unfortunately required γ as an input parameter; fortunately, the two authors of the two preceding methods teamed up to produce an *Ada*ptive Boosting method, AdaBoost, which does not require knowledge of γ, and preserved both the at majority structure and the $O(\ln(1/\varepsilon)/\gamma^2)$ size bound (Freund and Schapire, 1997).

This thesis studies AdaBoost and its immediate variants; the two main themes are as follows.

Duality. The coe cient values in the weighted majority function produced by AdaBoost are the solution to a convex optimization problem; the dual optimization is a maximum entropy problem over a space of weightings on individual books. The primal player seeks to identify good books; the dual player seeks a distribution which decorrelates all the characteristics from the quality of the books (thus forcing $\gamma = 0$). How much does duality dictate these basic behavior of the methods?

Data-dependence. The scalar γ is a function of data; speci cally, it is a function of the set of of books, and the set of chosen characteristics. The condition $\gamma > 0$ implies the size bound $O(\ln(1/\varepsilon)/\gamma^2)$ stated above. What other quantities and structures dictate the behavior of the algorithm, for example when $\gamma = 0$? Can these objects be derived from duality structure?

The purpose of this thesis is to study the duality structure and data-dependent structure of the optimization problems corresponding to AdaBoost variants, and thereafter to apply this lens to the algorithms themselves, deriving both numerical and statistical guarantees.

This introductory chapter is primarily devoted to establishing a language with which to discuss AdaBoost variants (Section 1.2 through Section 1.7, with additional technical development

in Appendix 1.A through Appendix 1.D). A summary of the main results of the thesis, as well as a more detailed overview of the remaining chapters, appears in Section 1.8.

1.2 The Hypothesis Class H

Let X denote an abstract input space (abstract room of abstract books). The aforementioned characteristics or correlates are modeled as a set of functions H, where each $h \in H$ has domain X and bounded range $[-1;+1]$, meaning $h : X \to [-1;+1]$. At times, it will be necessary to constrain every $h \in H$ to be binary, meaning $h : X \to \{-1;+1\}$.

Throughout this thesis, μ will denote a weighting over H, which leads to a map of the form

$$x \mapsto \sum_{h \in H} h(x)\ \mu(h):$$

By the boundedness of $h \in H$, this map also has bounded range, specifically $[-\|\mu\|_1; +\|\mu\|_1]$.

In order to more conveniently talk about this map, some further notation is beneficial. Firstly, let $\#$ denote counting measure over H, whereby the space $L^1(\#)$ is simply all weightings of H which are absolutely summable; that is to say, $\mu \in L^1(\#)$ implies

$$\|\mu\|_1 = \sum_{h \in H} |\mu(h)| < 1 \qquad \text{and} \qquad \sup_{x \in X} \sum_{h \in H} h(x)\ \mu(h) \leq \sup_{x \in X,\, h \in H} |h(x)|\, \|\mu\|_1 < 1\ ;$$

whereby the second property provides that the map $x \mapsto \sum_{h \in H} h(x)\ \mu(h)$ is well-defined whenever $\mu \in L^1(\#)$.

The algorithms considered here will only work with choices $\mu \in L^1(\#)$ which have finitely many nonzero coordinates. The reason for working with the larger space $L^1(\#)$ is that it is a complete normed vector space, which is essential to many continuity properties, which moreover are essential to many duality arguments. To this end, define the simpler notation $L^1 := L^1(\#)$.

As an additional convenience, let H be a linear operator over L^1 defined as

$$(H\ \mu)_x := (H\ \mu)(x) := \sum_{h \in H} h(x)\ \mu(h) = \int_H h(x)\ \mu(h)\, d\#(h);$$

where the preceding development established this sum is well-defined whenever $\mu \in L^1$. Note that this expression could have been defined directly over μ via some theory of integration over Banach spaces, for instance the Bochner integral, but pulling in that machinery here is unnecessary. For

further properties of H, please see Section 1.A.

In some of the analysis, namely in Chapters 2 and 3, both H and X will be finite, whereby H is simply a matrix. But either way, H is a linear operator, and Hf is well-defined for $f \in \mathcal{F}$.

1.3 Primal Optimization Problem

Identifying good and bad books is a classification problem; the boosting methods here will choose a set of coefficients $\lambda \in \mathcal{F}$, and rather than outputting $(H\lambda)_x$ for a book x, this value will be thresholded, meaning the final output is

$$x \mapsto 2 \cdot 1[(H\lambda)_x \geq 0] - 1,$$

which is 1 when $(H\lambda)_x \geq 0$, and -1 otherwise.

It is desirable for this map to output the correct value for every $x \in X$. In order to discuss correctness, the relevant space is now $X \times \{-1, +1\}$, the set $\{-1, +1\}$ distinguishing good and bad books. Correspondingly, let μ denote a probability measure over $X \times \{-1, +1\}$; at times, an i.i.d. sample $\{(x_i, y_i)\}_{i=1}^m$ from μ of size m will be given, and the corresponding empirical measure will be $b\mu$; many results hold for either measure, and thus the measure variable will often be employed. The σ-algebra is always the Borel σ-algebra; this is largely irrelevant, with the exception of material discussing how to approximate arbitrary (Borel!) measurable functions. Additionally, let μ_X denote the marginal distribution over X, with corresponding conditional probability $\Pr[y = +1 \mid x]$ (analogously define μ_X, $b\mu_X$, etc.).

With this measure notation in place, the goal of the algorithm is to have as few errors as possible over μ; that is, define the risk R of a measurable map $f : X \to \mathbb{R}$ as

$$R(f) := \int 1[(f(x) < 0 \wedge y = +1) \vee (f(x) \geq 0 \wedge y = -1)] \, d\mu(x)$$

(with $\hat{R}$ defined analogously with $b\mu$ replacing μ), and thus the goal is to approximately solve the optimization problem

$$\inf\{R(H\lambda) : \lambda \in \mathcal{F}\}. \tag{1.3.1}$$

There are two difficulties here. The first is that the algorithm will not have access to μ, but rather just an i.i.d. draw $\{(x_i, y_i)\}_{i=1}^m$ from μ. To overcome this, the empirical measure $b\mu$ corresponding

to this draw is used instead, meaning $\hat{R}$ is minimized rather than R. Of course, replacing μ with b is a dangerous proposition, and occupies much of this thesis. The main issue, as the problem is currently stated, is to constrain the complexity of H, or rather its convex hull; namely, it is essential to establish the convergence of these scores for h to those of μ uniformly over the class of functions $x \mapsto (H\ h)_x$.

The second issue is that this optimization problem is not tractable in the general case; for example, if the in mal error rate is 1% (because, say, H is too small, or simply the conditional probability $\Pr[\,Y = +1 \mid X\,]$ is not 0 or 1 almost everywhere), it is NP-Hard to solve this problem with fewer than 49% errors (Guruswami and Raghavendra, 2006). This problem will be overcome by replacing the in mand with a convex surrogate, developed as follows.

First, for added convenience, de ne another linear operator A, whereby $A\ h$ is now a function of both $x \in X$ and $y \in Y$:

$$(A\ h)(x,y) := (A\ h)_{x,y} := y(H\ h)_x = \sum_{h \in H} y h(x)\ \alpha(h);$$

note that this quantity is well-de ned for $h \in \{\hat{h}, b\}$ exactly for the same reasons that $H\ h$ is well-de ned. Now let ℓ be an nondecreasing convex function with $\ell(0) \geq 0$ and in particular $\ell(0) > 0$; for every $(x,y) \in X \times \{-1,+1\}$,

$$((H\ h)_x < 0 \wedge y = +1) \vee ((H\ h)_x \geq 0 \wedge y = -1) \implies (A\ h)_{x,y} \geq 0,$$

whereby

$$\mathbf{1}[((H\ h)_x < 0 \wedge y = +1) \vee ((H\ h)_x \geq 0 \wedge y = -1)] \leq \frac{\ell((A\ h)_{x,y})}{\ell(0)}. \tag{1.3.2}$$

This in turn motivates approximate minimization of the convex surrogate

$$\inf_{h \in \{\hat{h}, b\}} \int \ell((A\ h)_{x,y})\, d\mu(x,y) : \mu \in \Delta, \tag{1.3.3}$$

where Δ is either $\hat{\Delta}$ or Δ_b.

Of course, the inequality in eq. (1.3.2) is one-sided; ideally, approximate solutions to eq. (1.3.3) will also be approximate optima to the original (potentially nonconvex) problem in

eq. (1.3.1) Establishing a stronger relationship of this form is indeed possible, although it depends on a number of properties of ℓ and H (Zhang, 2004, Bartlett et al., 2006), and will be a key aspect of Chapter 6. For a hint as to why such a result is possible, note that if ℓ is differentiable and increasing at the origin (which is in fact a necessary condition (Bartlett et al., 2006, Theorem 2)), then it is nearly linear in some tiny open ball around the origin; as such, minimizing it can be expected to push incorrect predictions in the right direction, without hurting correct predictions too much.

The approach of AdaBoost (followed in this thesis) is to minimize the convex surrogate in eq. (1.3.3). The statistical issues with this optimization are more severe than with the non-convex formulation in eq. (1.3.1), since the function $R_\ell d$, unlike R, is sensitive to the norm $\|\cdot\|_1$; in particular, in order to achieve good statistical properties, it will sometimes be desirable to ensure that $\|\cdot\|_1$ is small.

Before adjourning this section, a few more remarks on these convex losses are appropriate, which leads to additional notation. First, the main losses considered throughout are

$$\begin{aligned}
\ell_{\exp}(x) &:= \exp(-x) && \text{exponential loss,}\\
\ell_{\log}(x) &:= \ln(1+\exp(-x)) && \text{logistic loss,}\\
\ell_{\text{hinge}}(x) &:= \max\{0, 1-x\} && \text{hinge loss,}\\
\ell_{\text{russ}}(x) &:= 0.5(x-1)^2 \mathbf{1}[-1 < x < 0] + (-x+0.5)\mathbf{1}[0 \le x] && \text{smoothed hinge loss}
\end{aligned}$$

The exponential loss $\ell_{\exp}$ was the original choice of AdaBoost. The logistic loss is popular in practice but less well understood theoretically, and therefore will receive preference in this thesis (in particular, some results in Chapter 6 hold for Lipschitz losses like $\ell_{\log}$, but not for the original choice $\ell_{\exp}$). The hinge loss ℓ_{hinge} is popular in machine learning, and an interesting counterpart to the preceding two because it is neither strictly convex nor differentiable; however, the only chapter to allow the hinge loss throughout is Chapter 5. Lastly, the smoothed hinge loss ℓ_{russ} is differentiable, but still not strictly convex, which will allow it to be used in a few places in Chapter 6 where ℓ_{hinge} is disallowed; this interesting loss is a rescaling of one implicit in a proof by Impagliazzo (1995, Proof of Lemma 1), and a translated version was also discussed by Zhang (2004, Section 3.6).

Finally, a few more notational conveniences. For any $f \in L^1(\mu)$ and $f \in L^1(b\mu)$, define the

shorthands $L(f) := \int \ell(f(x,y))\,d\pi(x,y) = \int \ell(f)\,d\pi$ and $\hat{L}(f) := \int \ell(f(x,y))\,d\hat{b}(x,y) = \int \ell(f)\,d\hat{b}$, where these definitions also demonstrate the practice of dropping integration variables. As such, a succinct way to write the central optimization problem from eq. (1.3.3), now specialized for $\hat{b}$, is

$$\inf\left\{ \hat{L}(A) : A \in \mathcal{A} \right\}. \tag{1.3.4}$$

Whenever L or $\hat{L}$ are passed a set S, let $L(S)$ and $\hat{L}(S)$ define minimization over that set, meaning for instance

$$L(A) := \inf\{L(A) : A \in \mathcal{A}\} \quad \text{and} \quad \hat{L}(A) := \inf\{\hat{L}(A) : A \in \mathcal{A}\},$$

with corresponding shorthands $R(H)$ and $\hat{R}(H)$ appearing frequently. Lastly, let brackets provide shorthand definitions of sets, meaning for instance

$$[A \geq 0] = [(A)_{x,y} \geq 0] = \{(x,y) \in X \times \{-1,+1\} : (A)_{x,y} \geq 0\}.$$

1.4 Dual Optimization Problem

The primal problem, presented in eq. (1.3.3) and eq. (1.3.4), is concerned with coefficients $A \in \mathcal{A}$ over the hypothesis space H. On the other hand, the dual is concerned with weightings over the instance space. To start, note the following *weak duality* result, which is proved in Section 1.B.

Proposition 1.4.1. Let finite measure π, hypothesis class H, and convex loss $\ell : \mathbb{R} \to \mathbb{R}_+$ with $\lim_{z \to -1} \ell(z) = 0$ be given. Then

$$\inf\left\{ \int \ell(A)\,d\pi : A \in \mathcal{A} \right\} \geq \sup\left\{ -\int \ell^*(p)\,d\pi : p \in L^1(\pi);\ p \geq 0\ \pi\text{-a.e.};\ \forall A \in \mathcal{A},\ \int (A)p\,d\pi \leq 0 \right\},$$

where ℓ^* denotes the *Fenchel conjugate* of ℓ, discussed below.

It would be nice to replace the inequality with an equality; this in turn is provided by strong duality statements, presented in Chapter 2 and Chapter 4, which will additionally provide optimality guarantees. Even so, this weak duality statement already identifies a few key properties of the dual.

First note that the Fenchel conjugate of a function $\ell : \mathbb{R} \to \mathbb{R}$ is defined as

$$\ell^*(\alpha) = \sup_{x \in \mathbb{R}} \alpha x - \ell(x).$$

The conjugation operation is a cornerstone of convex analysis, and studying the conjugacy properties of ℓ and L is essential to the development here. As discussed below, the choices of ℓ most relevant here have a conjugate which resembles entropy.[1] General information on Fenchel conjugates can be found in any treatment of convex analysis, for instance in the finite dimensional setting by Rockafellar (1970), and Hiriart-Urruty and Lemaréchal (2001) and Borwein and Lewis (2000), and in the in finite dimensional setting by Zalinescu (2002) and Borwein and Zhu (2005).

When π is a discrete measure over a finite subset of $\mathcal{X} \times \{-1, +1\}$ (e.g., when π is b), strong duality follows with a fair bit of generality, as is proved in Chapter 2. On the other hand, in the general case of an arbitrary (Borel) measure over $\mathcal{X} \times \{-1, +1\}$, strong duality is only proved (in Chapter 4) when ℓ is additionally Lipschitz, which rules out the exponential loss used by AdaBoost. There are indeed a number of steps that break down for the exponential loss, and in particular the dual should be over a larger space than $L_1(\pi)$, though this thesis only provides hints to the correct choice; please see Chapter 4 for further discussion. Note that the restriction of ℓ Lipschitz was reasonable since this allows for analysis of the logistic loss, which as discussed previously has been somewhat neglected despite its empirical success.

In the case of the conjugacy theory of L (upon which the duality builds), it is the combination of H in finite and π in finite which leads to problems; if either π is finite (cf. Proposition 1.B.2) or H is finite (cf. lemma 5.A.2), then L is much more well behaved.

This dual captures the duality structure stated in Section 1.1; namely, the dual space $L_1(\pi)$ is over reweightings p of the instance space the instance space $\mathcal{X} \times \{-1, +1\}$, and moreover the dual feasible set is all reweightings which decorrelate all primal weightings, written symbolically here as $\int (A \cdot \alpha) p \, d\pi = 0$ for all $\alpha \in \mathcal{A}$.

The machine learning literature follows both the convention here of having losses nonde-

[1] "You should call it entropy, [...] no one really knows what entropy really is" | John von Neumann.

Figure 1.1. The exponential loss $\grave{}_{\exp}(z) = \exp(z)$, the logistic loss $\grave{}_{\log}(z) = \ln(1 + \exp(z))$, and their conjugates.

creasing (see also (Boucheron et al., 2005b)), and the mirrored convention of nonincreasing losses. The choice here was made so that gradients and gradient-like functions of primal quantities will directly lead to nonnegative weightings over the source distribution; in this way, the gradients and dual domain are more readily interpreted in the context presented in Section 1.1.

As mentioned above, a few more remarks on the structure of $\grave{}$ are pertinent. First, the exponential loss $\grave{}_{\exp}$ is conjugate to the Boltzmann-Shannon entropy (see Borwein and Lewis, 2000, closing commentary, Section 3.3), defined as

$$\grave{}^{*}_{\exp}(\alpha) = \begin{cases} \alpha \ln(\alpha) & \text{when } \alpha > 0; \\ 0 & \text{when } \alpha = 0; \\ \infty & \text{otherwise;} \end{cases}$$

whereas the logistic loss $\grave{}_{\log}$ is conjugate to the Fermi-Dirac entropy (see Borwein and Lewis, 2000, closing commentary, Section 3.3), defined as

$$\grave{}^{*}_{\log}(\alpha) = \begin{cases} (1-\alpha)\ln(1-\alpha) + \alpha \ln(\alpha) & \text{when } \alpha \in (0,1); \\ 0 & \text{when } \alpha \in \{0,1\}; \\ \infty & \text{otherwise.} \end{cases}$$

For a plot of these losses and their conjugates, please see Figure 1.1.

(a) Stylized primal problem. (b) Stylized dual problem.

Figure 1.2. Stylized primal and dual problems; see for instance the weak duality relation in Proposition 1.4.1. On the primal problem, the arrow signi es that the objective value may fail to ever curve back upward; correspondingly, not only may minimizers fail to exist, but moreover it may be necessary for a sequence of iterates approaching the optimal value to have unbounded norms. On the other hand, the dual problem possesses curvature (cf. Proposition 1.4.2). The green area signi es that the dual is a constrained optimization problem: in particular, it is maximizing this curved shape, subject to the constraint that all weak predictors are decorrelated from the target prediction value f 1; +1 g.

More generally, while the primary losses considered here do not have minima (they tend to zero), their conjugates have a bowl-like structure, attaining a minimum and increasing to either side of it.

Proposition 1.4.2. Suppose` : R ! R+ is di erentiable, strictly convex, and lim z! 1 `(z) = 0 . Then

$$\grave{}(\ ) = \begin{cases} 1 & \text{when } < 0; \\ 0 & \text{when } = 0; \\ (\ \grave{}(0); 0) & \text{when } 2 (0; \grave{}^{0}(0)); \\ \grave{}(0) & \text{when } = \grave{}^{0}(0); \\ (\ \grave{}(0); 1\ ] & \text{when } > \grave{}^{0}(0): \end{cases}$$

Correspondingly, note the stylized versions of the primal and dual optimization problems as depicted in Figure 1.2.

1.5 Algorithm

What is the best way to minimize the primal optimization problem in eq. (1.3.4)? Ideally, the method somehow fits with the computational model suggested in Section 1.1, namely that a good correlation variable $h \in H$ can be found, given a reweighting of the provided examples $\{(x_i, y_i)\}_{i=1}^m$.

The choice followed by AdaBoost, and continued here, is to use coordinate descent: each iteration updates a coefficient vector $\alpha \in \mathbb{R}^m$ by computing a gradient, and then updating the coordinate of α (identified by some $h \in H$) corresponding to a large entry in the gradient.

Concretely, coordinate descent is applied to the function $\hat{\mathcal{L}}_A$; as is developed in Proposition 1.B.2, the gradient of $\hat{\mathcal{L}}_A$ satisfies the familiar relation $\nabla(\hat{\mathcal{L}}_A)(\alpha) = A^\top \nabla\hat{\mathcal{L}}(A\alpha)$, but more importantly, it satisfies the pairing

$$\left\langle \nabla(\hat{\mathcal{L}}_A)(\alpha), e_h \right\rangle = \frac{1}{m}\sum_{i=1}^m y_i \ell'(-(A\alpha)_{x_i,y_i}) h(x_i); \tag{1.5.1}$$

where $e_h \in \mathbb{R}^H$ is the indicator vector for a single $h \in H$. Notice in particular that this precisely matches the computational model of interest here: this inner product is simply the correlations between h and the labels $\{y_i\}_{i=1}^m$, reweighted by $\nabla\hat{\mathcal{L}}(A\alpha)$. For each $h \in H$, supposing α has finite support (which it will within the algorithm), this expression is expressly computable (given, of course, some computable approximation of the reals and ℓ). As a technical note, the gradient used here is the Gâteaux derivative, which can handle an infinite coordinate space H, but crucially does not disagree with the usual notion when everything is finite; again, these details are presented together with basic convexity properties of $\hat{\mathcal{L}}$ in Proposition 1.B.2.

Coordinate descent methods typically choose the best coordinate. Since H may be infinite, a best coordinate may not exist; but even if it did, it is useful for computational reasons to relax this requirement. As such, the algorithm here takes in a parameter $\tau \in (0, 1)$, and each coordinate update chooses some $h_t \in H$ so that

$$\left|\nabla\hat{\mathcal{L}}(A\alpha_{t-1})^\top A e_{h_t}\right| \geq \sup_{h\in H}\left|\nabla\hat{\mathcal{L}}(A\alpha_{t-1})^\top A e_h\right| = \left\|\nabla\hat{\mathcal{L}}(A\alpha_{t-1})^\top A\right\|_1;$$

where the middle expression matches embeds eq. (1.5.1) from above (and thus fits the computational model), and the correspondence with $\|A^\top \nabla\hat{\mathcal{L}}(A\alpha_t)\|_1$ in the last expression is provided by

Lemma 1.A.5, which is again consistent with the nite case. Note that the absolute value means that a descent step will move either in direction $+h_t$ or $-h_t$; many treatments of AdaBoost omit this choice, but instead stipulate that H is closed under negation, meaning $h \in H \Leftrightarrow -h \in H$.

The full algorithm appears in Figure 1.3. It contains two choices which have not yet been pointed out.

1. There are two possible stopping criteria. One is based on the duality gap; unfortunately, as discussed Chapter 2, this choice is not computationally feasible, since it requires determining the kernel of A^∞, whereas the computational model stressed here does not even require A and H to be ever be stored in memory. The second choice is to simply stop after $\lceil m^a \rceil$ iterations, where $a \in (0, 1)$ is provided in advance. A popular choice in practice, which is not studied here, is to use cross-validation.

2. Three line search choices are allowed. When the unconstrained step exists, and H is binary and ℓ is the exponential loss, then this step recovers the original choice used in AdaBoost. The unconstrained step is the most di cult to analyze, since it may elect to take extremely large steps, potentially damaging the aforementioned suggestion that solutions with small norm are preferable. This di culty will be acute in Chapter 6.

   The second choice, or rather range of choices, is based upon computing a quadratic upper bound to the line search problem, and choosing its minimizer. In some cases, this step also has a simple expression, and does not lead to the di culties of the unconstrained step.

   Lastly, the Wolfe search is somewhere between these two; in a sense, it attempts to minimize the univariate line search problem, but subject to a constraint that disallows steps from being too large. The implementation of this method is similar to binary search, and does not need to know any properties of the functions it is minimizing (in the convex case). It takes two parameters, which are set to 1/2 and 1/3 in this thesis for simplicity.

   Details on all three line searches may be found in Section 1.C. Each provides nearly the same optimization guarantees, however, as discussed above, they vary in terms of the norms of coe cient vectors they lead to.

Routine. Boost .
Input. Loss ℓ and empirical measure $\hat{b}$ (granting m, $L\hat{b}$, $\nabla L\hat{b}$), hypothesis class H (granting A, $A^\top$), coordinate search parameter $\gamma \in (0, 1]$.
Output. Coefficient vectors $\{\alpha_i\}_{i=1}^t$.

1. Set $\alpha_0 := 0$.
2. For $t = 1, 2, \ldots$:
   (a) Decide whether to exit loop over t as follows.
      Option 1. Exit if $t \geq m^a$, for some provided $a \in (0, 1)$.
      Option 2. Exit if duality gap discussed in Remark 2.2.5 is upper bounded by some provided $\varepsilon > 0$.
   (b) Choose approximate best coordinate (weak learner) $h_t \in H$ satisfying
   $$|\nabla L\hat{b}(A\alpha_{t-1})^\top A e_{h_t}| \geq \sup_{h \in H} |\nabla L\hat{b}(A\alpha_{t-1})^\top A e_h| = \gamma \|\nabla L\hat{b}(A\alpha_{t-1})^\top A\|_1 .$$
   (c) Set descent direction $v_t \in \{-e_{h_t}\}$, whereby
   $$\|A^\top \nabla L\hat{b}(A\alpha_{t-1})\|_1 \cdot \nabla L\hat{b}(A\alpha_{t-1})^\top A v_t = -\|A^\top \nabla L\hat{b}(A\alpha_{t-1})\|_1 .$$
   (d) Set $\alpha_t^* := \arg\min_{\alpha \geq 0} L\hat{b}(A(\alpha_{t-1} + \alpha v_t))$, and choose a step α_t as follows.
      Option 1. If $\alpha_t^* < 1$, set $\alpha_t = \alpha_t^*$ (i.e., make an optimal unconstrained step).
      Option 2. If $L\hat{b}$ has Lipschitz gradients with parameter B_t as discussed in Section 1.C, choose any $\alpha_t \in [\gamma \cdot \nabla L\hat{b}(A\alpha_{t-1})^\top A v_t / B_t, \alpha_t^*]$.
      Option 3. Choose any α_t satisfying the Wolfe conditions (please see Section 1.C.2.)
   (e) Update $\alpha_t := \alpha_{t-1} + \alpha_t v_t$.
3. Return $\{\alpha_i\}_{i=1}^t$.

Figure 1.3. The boosting algorithm: coordinate descent applied to $L\hat{b} \circ A$. Note that AdaBoost (Freund and Schapire, 1997) is recovered by choosing the exponential loss $\ell = \ell_{\exp}$, exact unconstrained step sizes (option 1), and exact coordinate selection ($\gamma = 1$).

1.6 Other Approaches

Before descending into a labyrinth of analysis of this algorithm and optimization problem, it is a good time to consider other approaches to the same problem.

It should be stressed that tools from convex optimization typically work with objective functions which are strongly convex (lower bounded by a quadratic), or constrained. It is also common to assume the existence of minimizers, which is characterized in Chapter 2 as imposing quite a bit of curvature structure on the objective function. On the other hand, the original AdaBoost analysis of Freund and Schapire (1997), under the assumption > 0, is characterized in Chapter 2 as being, in a strong sense, diametrically opposed to the typical setting of convex optimization. These two distinct settings, and their interplay, is a recurring element in this thesis.

Back to task, here are some other ways to solve this problem.

More coordinates. Leave the objective function alone, but take more than one coordinate at a time. Without any further assumptions or algorithmic modi cations, it could be the case that all the chosen coordinates are similar, and therefore not much is gained over choosing a single coordinate (Long and Servedio, 2011). On the other hand, possibly with additional assumptions or an explicit choice of H as possessing some sort of independence between elements, there could be a way to still make good progress (Dasgupta and Long, 2003).

Subgradient Descent. Leave the objective function alone, suppose $\hat{e}$ is Lipschitz, assume H is nite, and simply execute subgradient descent. The analysis of subgradient descent typically assumes a minimizer, but an inspection of the standard proof reveals that no optimality properties of this minimizer are invoked, and the proof may be simply restated with respect to an arbitrary reference point. In particular, reworking the clean presentation by Vandenberghe (2013), it follows, for any $C > 0$, that

$$\hat{L}(A_{\text{best-of-}t}) \leq \inf_{k \, k_2 \leq C} \hat{L}(A_k) + \frac{CL}{\sqrt{t}};$$

where $_{\text{best-of-}t}$ is the element encountered with lowest value of $\hat{L}$ in the rst t iterations, and L is a Lipschitz parameter. This rate is slower than those presented in Chapter 3, and of course the computational model has been departed (the operator A, now simply a matrix, needs to be explicitly computed), however the method is simple and the comparison to a

solution with small norm is statistically pleasing.

Naive Regularization. Another choice is to choose some regularization weight > 0, and replace eq. (1.3.4) with

$$\min \left\{ \hat{L}(A\ ) + \ k\ k_1 : \ 2 \right\};$$

where the use of "min" is justified, meaning a minimizer always exists. From here, a variety of techniques may be employed to solve this problem; but of course, there is a question of how to choose the parameter .

Statistical properties of penalized boosting estimators were studied by Blanchard et al. (2003), Lugosi and Vayatis (2004). Since is is not entirely clear how to implement these algorithms, one simple choice achieving a similar effect is to explicitly curtail the step size in Boost , which was shown by Zhang and Yu (2005) to lead to good numerical and statistical rates. While this thesis does not close the question of how eq. (1.3.4) should best be penalized, it is interesting to point out that there are cases, as studied in Chapter 6, where excellent rates are achieved despite producing solutions with large norms.

Different Objective. Another choice is to simply rewrite the objective function entirely; for instance, Rätsch and Warmuth (2005), Shalev-Shwartz and Singer (2008) add both regularization and constraints. Some basic guarantees of this scheme have been provided, but neither this thesis, nor other existing literature, comprehensively identify the respective benefits or disadvantages of the scheme considered here versus these constrained schemes.

1.7 Basic Structures

To give a taste of future material and to clarify a few ambiguous claims, this section develops a few basic quantities relevant throughout the analysis.

The quantity , discussed in Section 1.1, is typically called the *weak learning rate*, and the condition > 0 is called the *weak learning assumption*. In the original discussion, it was not clear whether this value is measured over or b, thus consider the following definition.

Definition 1.7.1 (Schapire and Freund (2012, Chapter 2)) A class H is *weakly PAC-learnable* with rate if for any measure over $X \quad f \ 1; +1 g$, there exists $h \ 2 \ H$ with $\int yh(x)d \ (x; y) \quad$.

Additionally, class H and empirical measure $$ are empirically weakly learnable with rate $( )$ if there exists $h \in H$ so that $\int yh(x)p(x;y)d(x;y) \quad ( )$ for every reweighting p of measure $$.

As discussed previously, the assumption $(b) > 0$ (which is implied by > 0) is greatly beneficial to optimization, with AdaBoost needing only $O(\ln(1/)/ (b)^2)$ iterations to achieve accuracy > 0.

What are some ways for the weak learning assumption $(b) > 0$ to fail? An easy case is that the sample $\{(x_i;y_i)\}_{i=1}^m$ has $x_i = x_j$ for some $i \neq j$, but $y_i \neq y_j$: the dual player can simply place positive probability on these two examples, and zero probability elsewhere. Additionally, it is possible that the sample is separable but each element of H only breaks even by a constant number of examples, meaning $(b) = O(1/m)$, which will be momentarily identified as inadequate for statistical guarantees based (b).

Remark 1.7.2. The largest achievable empirical weak learning rate (b) is equivalent to a measure of the separability of an instance, called the maximum l_1 margin (Shalev-Shwartz and Singer, 2008). Moreover, AdaBoost, with a number of step size and loss function choices, is able to achieve nearly maximal l_1 margins in a small number of iterations (Telgarsky, 2013). When (b) is positive and not too small, for instance when $\sqrt{V(H)/m}/ (b) \to 0$ as $m \to 1$ (where $V(H)$, the VC dimension of H, measures the complexity of H), AdaBoost exhibits good generalization performance (Schapire et al., 1997).

As the analysis in this thesis is primarily concerned with worst case guarantees, this route of margin assumptions and analysis is not explicitly considered, though ideas from it appear in various places. Since large or even positive margins can not be guaranteed, thus guarantees on b and L are more generally applicable.

It is worth mentioning, however, when there is prior knowledge on the structure of the prediction problem encoded in $$, it may be possible to engineer a hypothesis class H which ensures large margins, and moreover, the engineering effort involved in implementing AdaBoost variants is typically largely spent in this process.

As a consequence, it is arguably AdaBoost's greatest strength that it possesses both an excellent set of margin-based bounds under favorable values of (b), and a strong set of fallback guarantees based on loss minimization properties.

In the previous section, the question of L_b possessing minimizers was raised; how does

this relate to the weak learning assumption? As discussed previously, and as will be proved in Chapter 3 for choices of ℓ which include $\ell_{\exp}$ and $\ell_{\log}$, this assumption implies that $\inf(A_t)$ decreases exponentially quickly, which in particular means $\inf(A) = 0$. On the other hand, since $\ell_{\exp} > 0$ and $\ell_{\log} > 0$, it follows that $\inf(A) > 0$ for either loss and any $\in$: in particular, the empirical weak learning assumption $(b) > 0$ and the existence of minimizers never coincide. But what is a more complete characterization of the objective function?

To this end, another basic structure related to L and $\inf$ is the hard core, defined as follows.

Definition 1.7.3. Given finite hypothesis class H and probability measure π, let $D(H, \pi)$ denote reweightings of π which decorrelate every regressor $h \in H$; that is,

$$D(H, \pi) := \left\{ p \in L^1(\pi) : p \geq 0 \ \pi\text{-a.e.}, \forall h \in H, \int (A_h) p \, d\pi = 0 \right\}.$$

Correspondingly, $S_D(H, \pi)$ tracks the supports of these weightings:

$$S_D(H, \pi) := \{ [p > 0] : p \in D(H, \pi) \}.$$

A hard core $C \subseteq X \times \{-1, +1\}$ for (H, π) is a maximal element of $S_D(H, \pi)$; that is,

$$C \in S_D(H, \pi) \quad \text{and} \quad \forall C' \in S_D(H, \pi) \ (C' \setminus C) = 0.$$

("Maximal", in the presence of measures, will always mean up to sets of measure zero.)

In words, $S_D(H, \pi)$ is a collection of sets for which there exists reweightings $p \in L^1(\pi)$ which decorrelate every function $h \in H$ from the prediction problem encoded in π, meaning $\int (A_h) p \, d\pi = \int (A_h)_{x,y} p(x, y) \, d\pi(x, y) = 0$. And amongst the elements of $S_D(H, \pi)$, a hard core C is maximal in a measure-theoretic sense, meaning no element of $S_D(H, \pi)$ exceeds C by a set of positive measure.

Crucially, hard cores always exist.

Theorem 1.7.4. Every finite hypothesis class H and probability measure π admits a hard core, and any two hard cores disagree only on a set of measure zero.

To prove this, note first that $S_D(H, \pi)$ is never empty—it will at least contain the empty set (with corresponding weighting $p = 0$). Furthermore, $S_D(H, \pi)$ is closed under countable

unions; but simply unioning together all of $S_D(H;)$ will not produce a hard core, since the corresponding object need not be measurable; the full proof is in Section 1.D.

The hard core provides a handle on the relationship between weak learnability and minimizers. Speci cally, in the nite sample optimization guarantees of Chapters 2 and 3, any nite sample can be split uniquely into two pieces, where one piece satis es the weak learning assumption, and the other has minimizers, and moreover considering these two pieces separately leads to a general analysis. Thereafter, Chapter 5 uses hard cores to similarly split the statistical deviations into two pieces, and thereafter control each piece separately.

1.8 Analysis Overview

The main byproducts of this thesis are a number of worst case and high probability guarantees for AdaBoost variants as de ned byBoost in Figure 1.3; in particular, although it is a fruitful and important topic, this thesis does not focus on scenarios where $(b) > 0$ is favorably non-small as discussed in Remark 1.7.2. Concretely, the rst set of guarantees are of numerical convergence, namely showing the ability ofBoost to minimize L^b or in some cases R^b (Chapter 3 and Chapter 6). The second set of guarantees are of statistical convergence, proving that the output sequence of iterates also minimizes L or in some cases R (Chapter 5 and Chapter 6), which in turn can lead to statements of statistical consistency (Chapter 6).

In order to systematically develop these byproducts, this thesis devotes signi cant e ort to identifying the structure of the primal and dual optimization problems identi ed in Section 1.3 and Section 1.4, and moreover their impact on the behavior of the algorithm. This investigation, initiated in this introductory chapter (particularly in the immediately preceding Section 1.7), is the focus of Chapter 2 and Chapter 4, but can also be found throughout the other chapters as well.

In more detail, the complete layout of this thesis is as follows. Note that most proofs are relegated to appendices closing each chapter.

Chapter 2. This chapter and the following focus on nite samples, meaning the empirical objective L^b; additionally, most results assume that H has some form of bounded complexity, whereby the operators A and H, restricted to the sample, may be written as nite dimensional matrices.

This chapter first proves a strong duality result, which is used to characterize hard cores over finite samples. Additionally, this chapter develops an analog to (b) which is never zero, and thus is a data-dependent quantity which characterizes the behavior of the algorithm without any assumptions.

Chapter 3. Leveraging the structures in Chapter 2, optimization guarantees follow. In particular, when either $(b) > 0$ (the hard core is empty) or minimizers exist (the hard core is the entire sample), the number of iterations to reach accuracy > 0 is $O(\ln(1/\epsilon))$. For general instances, the problem may be broken into the hard core and its complement, analyzed using the preceding tools on both pieces, and then stitched together to achieve a rate $\Theta(1/\epsilon)$, with a matching lower bound for the logistic loss $\ell_{\log}$.

Chapter 4. This next chapter starts the second part of the thesis, which focuses on statistical questions. Chapter 4 is a gentle introduction, presenting a strong duality analog to Proposition 1.4.1, and identifying where difficulties begin to arise when the measure is not assumed to correspond to a finite sample.

Chapter 5. The first set of statistical guarantees are against an arbitrary measure (with potentially infinite support), but still assuming H is finite. While this is not quite the desired level of generality, this family of results have two points of value. First, the results are stronger than those presented later in Chapter 6 for H infinite: more loss functions are handled, the deviation bounds are tighter, and less demands are made on the minimization algorithm. Second, note that this guarantee may be invoked by starting with an infinite H, and then feeding an appropriate finitization to an algorithm; in a sense, this indicates that an effective penalization scheme for boosting is to determine how much data is available, and then appropriately curtail the complexity of H before invoking the algorithm.

The method of proof here is to develop the structure of hard cores more carefully in the case of a general measure (and not simply an empirical measure corresponding to a finite sample as in Chapter 2).

Chapter 6. Lastly, losses which are Lipschitz (plus a few other regularity conditions) are handled in the case that H is infinite; amongst the four losses presented in Section 1.3, only the logistic loss $\ell_{\log}$ survives, though some of the analysis also holds for ℓ_{russ}. The analysis here

considers two cases: either the in mal surrogate risk over the distribution is zero, or it is positive. When it is zero ($L(A) = 0$), the convergence rate is fast, and moreover based on an analog to which is statistically stable and in a sense better adapted to Lipschitz losses. On the other hand, when the in mal surrogate risk is positive ($L(A) > 0$), it is possible to exhibit distribution-dependent controls on the norms of the output predictors, which again leads to statistical controls.

Appendix 1.A Banach Spaces and Linear Operators

The purpose of this section is to identify basic properties of the Banach spaces and linear operators relevant in this thesis, which are fundamental to the behavior of convex duality theory (since the dual is nothing but a set of linear functions and a topology that renders them continuous); in particular, this section will provide the de nition of the dual pairing $h\,;\,i$, which is used in the de nition of Fenchel conjugate.

H and A are mappings which produce bounded functions; the bulk of the analysis, however, considers them as producing functions over $L^1(\ ^X)$ and $L^1(\ )$ as follows (where is a probability distribution over $X \quad f \quad 1; +1 g$).

Lemma 1.A.1. Let be a probability measure over $X \quad f \quad 1; +1 g$, and let ^{X} denote the marginal distribution over X.

1. The de nition of H and A is valid for arbitrary weightings 2 ; in particular, $\mathrm{supp}(\ )$ is countable, and

$$(H\ )_x = \int h(x)\ (h)d\ (h) = \sum_{h2\,\mathrm{supp}(\ )} (H\ )_x\ (h);$$

$$(A\ )_{x;y} = \int y\ (H\ )_x\ (h)d\ (h) = y \sum_{h2\,\mathrm{supp}(\ )} (H\ )_x\ (h);$$

2. H and A are linear operators.

3. $H : \ ! \ L^1(\ ^X)$ and $A : \ ! \ L^1(\ )$ are continuous linear operators (with unit norm).

Proof. If 2 , then $\int j\ (h)jd\ (h)j < 1$, and since is a counting measure, it follows that $\mathrm{supp}(\ )$ is countable. Furthermore, for any $x; y; h$, $jh(x)j \quad j \quad yh(x)j \quad 1$, and thus the rescalings

$h(x)$ and $yh(x)$) are both in , and in particular

$$(H\ )_x = \int h(x)\ (h)d\ (h) = \sum_{h2\,\mathrm{supp}(\ )} (H\ )_x\ (h);$$

and similarly for A.

It follows by de nition (and another check for integrability) that $(\ H(a\ _1 + b\ _2))_x = a(H\ _1)_x + b(H\ _2)_x$, and thus H is a linear operator; the proof for A is the same.

Lastly, H is continuous with unit norm, since boundedness of each$h$ combined with ^{X} being a probability measure gives

$$\begin{aligned}
\sup\{k H\ k : \ 2\ ; k\ k_1\ \ 1\} &= \sup \int (H\ )d\ ^X : \ 2\ ; k\ k_1\ \ 1 \\
&\sup\left\{ \int \sum_{h2\,\mathrm{supp}(\ )} h(x)\ (h)d\ ^X(x) : \ 2\ ; k\ k_1\ \ 1 \right\} \\
&\sup\left\{ \sup_{x;h} |h(x)| \sum_{h2\,\mathrm{supp}(\ )} (h) : \ 2\ ; k\ k_1\ \ 1 \right\} \\
&= 1:
\end{aligned}$$

(The proof for A is the same, since$y\ 2\ f\ \ 1; +1\,g$ implies $jyh(x)j = jh(x)j$.) □

Note, H and A may also be de ned as Bochner (or similar) integrals.

Next, to develop the adjoint of $A^>$, relevant dual spaces need to be established (the adjoint of $H^>$ does not appear, but is similar).

Lemma 1.A.2. If is a probability measure over$X$ $f\ \ 1; +1\,g$, then $L^1(\ )$ is isometrically isomorphic to $L^1(\ )$, and in particular for every $Q\ 2\ L^1(\ )$ there exists$q\ 2\ L^1(\ )$ so that $Q(f) = \int qfd$ for every $f\ 2\ L^1(\ )$. Similarly, recalling $= L^1(\ )$ where is counting measure over some class$H$, the dual $L^1(\ )$ is isometrically isomorphic to $L^1(\ )$, and once again elements of $L^1(\ )$ can be written as integrals over with an element of$L^1(\ )$.

Proof. The rst relationship follows since is a probability measure and thus - nite (Folland, 1999, Theorem 6.15), and the second is a general property of counting measures (even though the cardinality of H may preclude from being - nite) (Folland, 1999, Exercises 3.15 and 6.25). □

This thesis always identi es the above dual spaces by the provided isometric isomorphism,

a fact which will be crucial in the convex duality theory of L (cf. Lemma 4.1.1). The following definition clarifies this impact.

Definition 1.A.3. Let X and X^* be a Banach space and a dual space. Given $x \in X$ and $x^* \in X^*$, define the dual pairing $\langle x, x^* \rangle = x^*(x)$ (Rudin, 1973, Chapter 4). When X is identified via isometric isomorphism, this dual pairing provides the means to compute these linear functions. In particular, in light of Lemma 1.A.2:

Given σ and $q \in L^1(\sigma)$, $\langle p, q \rangle = \int p(h)q(h)d\sigma(h)$.

Given measure β over $X \times \{-1, +1\}$, as well as $f \in L^1(\beta)$ and $p \in L^1(\beta)$, $\langle f, p \rangle = \int f(x,y)p(x,y)d\beta(x,y)$.

Given empirical measure $\hat{\beta}$ corresponding to a finite sample of size m, as well as $f \in L^1(\hat{\beta})$ and $p \in L^1(\hat{\beta})$, $\langle f, p \rangle = \sum_{i=1}^{m} f(x_i, y_i)p(x_i, y_i) = m \int f p d\hat{\beta}$.

Remark 1.A.4. Note that the dual pairings with respect to $L^1(\sigma)$ and $L^1(\hat{\beta})$ are not consistent, which for instance causes the artifact of an "m" appearing in the dual form over $L^{\hat{\beta}}$ as compared with L (concretely compare Theorem 2.2.1 to Proposition 1.4.1 and lemma 4.1.1. By making the dual pairing over $L^1(\hat{\beta})$ agree with the standard scalar product, however, duality and in particular gradient manipulations of $L^1(\hat{\beta})$ are consistent with the standard forms over finite dimensional Euclidean spaces, which hopefully improves readability of any such manipulations, particularly those in Chapters 2 and 3, and moreover allows the application of results from the literature which assumed the dual pairing is simply the scalar product.

To avoid confusion, dual pairing notation will often be avoided, but notice that it is at the heart of the convex duality relations, since the Fenchel conjugate to a function $g : X \to \mathbb{R}$ is defined as

$$g^*(x^*) = \sup_{x \in X} \langle x, x^* \rangle - g(x).$$

Turning back to H and A, the adjoint A^* has the following structure.

Lemma 1.A.5. Let probability measure σ over $X \times \{-1, +1\}$ and any H be given.

1. Considering A as a linear operator from σ to $L^1(\sigma)$, its adjoint $A^* : \sigma \to L^1(\sigma)$ is the unique continuous linear operator satisfying $(A^* p)(\sigma) = \int (A \sigma) p$, where $p \in L^1(\sigma)$ and $\sigma \in \sigma$ (and dual spaces have been identified via isomorphism as in Lemma 1.A.2).

2. Again identifying $A^{>}p$ for $p \in L^1(\ )$ with an element of $L^1(\ )$,

$$\begin{aligned} \|A^{>}p\|_1 &= \sup \left\{ \int yh(x)p(x;y)d\ (x;y) : h \in H \right\} \\ &= \sup \left\{ \int (A\ )_{x;y}\, p(x;y)d\ (x;y) : \ \in \ ; \|\ \|_1 \ \ 1 \right\} : \end{aligned}$$

3. The map $p \mapsto \|A^{>}p\|_1$ is a convex function over $L^1(\ )$, and is lower semi-continuous in the weak* topology (i.e., the weak topology induced on $L^1(\ )$ by $L^1(\ )$).

Proof. 1. Recall by Lemma 1.A.1 that A is a continuous linear operator; the basic properties of $A^{>}$ follow by properties of adjoints of continuous linear operators (Rudin, 1973, Theorem 4.10) combined with the isometric isomorphism of the relevant dual spaces as provided by Lemma 1.A.2.

2. Let $p \in L^1(\ )$ be given. Since the isometric isomorphism provided by Lemma 1.A.2 allows $A^{>}p$ to be identified with an element of $L^1(\ )$, the operator norm of $A^{>}p$ is simply the $L^1(\ )$ norm of the element it has been identified with by the isomorphism. Since is a counting measure, letting $e_h \in$ be an indicator function for a single $h \in H$ (it is 1 on h and 0 elsewhere), using the above adjoint relation $(A^{>}p)(e_h) = \int (Ae_h)p$, and using the definition of norms on $L^1(\ )$,

$$\begin{aligned} \|A^{>}p\|_1 &= \inf \{ a \ \ 0 : (f h \in H : |(A^{>}p)(e_h)| > a g) = 0 \\ &= \inf \left\{ a \ \ 0 : (f h \in H : \int yh(x)p(x;y)d\ (x;y) > a g) = 0 \right. \\ &= \inf \left\{ a \ \ 0 : 8h \in H ; \int yh(x)p(x;y)d\ (x;y) \ \ a \right. \\ &= \sup_{h \in H} \int yh(x)p(x;y)d\ (x;y) \ ; \end{aligned}$$

where the last equality can established by noting the domain of the infimum includes all $a \ \ 0$ satisfying $a \ \ \sup_{h \in H} \int yh(x)p(x;y)d\ (x;y)$, but no values satisfying $a < \sup_{h \in H} \int yh(x)p(x;y)d\ (x;y)$.

Next, to show

$$\sup\left\{\int yh(x)p(x;y)d\ (x;y)\ :\ h\ 2\ H\right\}$$
$$= \sup\left\{(A\ )_{x;y}\int p(x;y)d\ (x;y):\quad 2\quad ;k\ k_1\quad 1\quad ;\right.$$

one direction is immediate, since positive and negative copies e_h of the indicator elements satisfy $e_h\ 2$ and $k\quad e_h k_1 = 1$. For the other direction, let $\quad > 0$ be arbitrary, and choose any $\quad 2$ which is within $\quad$ of the supremum on the right side of the display. Then, since $jsupp(\ )j$ is countable (via Lemma 1.A.1), and since $k\quad k_1\quad 1$ implies $kA\quad k_1\quad 1$, the dominated convergence theorem (Folland, 1999, Theorem 2.25 (summation form)) may be applied (with dominating function 1), and

$$\begin{aligned}
&\sup\left\{(A\ )_{x;y}\int p(x;y)d\ (x;y):\quad 2\quad ;k\ k_1\quad 1\right.\\
&\quad + \int (A\quad)_{x;y}\int p(x;y)d\ (x;y)\\
&= \quad + \int \sum_{h\ 2\ supp(\quad)} yh(x)\quad (h)p(x;y)d\ (x;y)\\
&= \quad + \sum_{h\ 2\ supp(\quad)} (h)\quad \int yh(x)p(x;y)d\ (x;y)\\
&\quad + k\quad k_1 \sup\left\{\int yh(x)p(x;y)d\ (x;y)\ :\ h\ 2\ H\quad ;\right.
\end{aligned}$$

since $k\quad k_1\quad 1$, and since $\quad > 0$ was arbitrary, the result follows.

3. For the last part, de ne a convex indicator over $L^1(\ )$ as

$$(f\ ) = \begin{cases} 0 & \text{when } f\ 2\ f A\ :\quad 2\quad ;k\ k_1\quad 1g; \\ 1 & \text{otherwise:} \end{cases}$$

(Note that $\quad$ is not necessarily lower semi-continuous over $L^1(\ )$, since as discussed shortly in Lemma 1.A.6, the subspace $A\ $ might not be closed.) The conjugate of $\quad$ is, for any

$p \in L^1(\ )$,

$$\begin{aligned}(p) &= \sup \int f p : 9 \ 2 \ ; k \ k_1 \ 1 \ f = A \\ &= \sup \int (A \ )p : \ 2 \ ; k \ k_1 \ 1 \\ &= k A^{>} p k_1 \ ;\end{aligned}$$

where the last step used the earlier equalities for $A^{>}$. Since $p \mapsto k A^{>} p k_1$ is the conjugate of a convex function, it is lower semi-continuous in the weak* topology (Zalinescu, 2002, Theorem 2.3.1(i)).

□

Lastly, note the following properties of the sets H and A.

Lemma 1.A.6. Let any H and any probability measure over $X \ f \ 1; +1g$ with marginal ^{X} over X be given. Then H and A are subspaces, but it is possible that neither is closed in its respective $L^1(\ ^X)$ and $L^1(\ )$ topology (indeed, Example 6.3.4 provides the counterexample).

Proof. Since is a Banach space and H and A are linear operators, it follows that H and A are subspaces.

For the lack of closure, consider the setting of Example 6.3.4, and in particular building a sequence of functions $f f_k g_{k=1}^1$ which are a combination of k thresholds, and predict correctly on the last k intervals. This sequence has a limit point in $L^1(\ ^X)$ (in particular, it is a countable sum of indicators over intervals), but no such function is in H, which is therefore not closed in $L^1(\ ^X)$. To obtain a similar result for A, de ne $g_k(x; y) := f_k(x)$. □

Appendix 1.B Convexity Properties of ℓ and L_b

First, the following basic properties of ℓ are used throughout the thesis.

Lemma 1.B.1. Suppose $\ell : R \to R_+$ is convex with $\lim_{z \to 1} \ \ell(z) = 0$.

1. ℓ is lower semi-continuous, whereby ℓ is convex lower semi-continuous, and $\ell = \ell$.
2. $\ell(\ ) = 1$ for < 0, and $\ell(0) = 0$.

3. Let $:= \sup_{x \neq y} |\grave{}(x) - \grave{}(y)| / |x - y|$ denote the tightest Lipschitz constant for $\grave{}$. If < 1, then $\grave{}(\ ) = 1$ when $>$, and $\grave{}(\ ) < 1$ when $2 [0;]$. If $= 1$, then $\grave{}$ is nite along $[0; 1)$.

4. If $g 2 @\grave{}(0)$ is any subgradient of $\grave{}$ at the origin and $\grave{}(0) > 0$, then $\grave{}(\ ) < 0$ for $2 (0; g)$, and $\grave{}$ attains its minimum value $\grave{}(0)$ at g.

5. If $\grave{}$ is di erentiable and strictly convex, then $\grave{}$ is strictly convex and continuously di erentiable over the interior of its domain $\mathrm{dom}(\grave{}) = \{ 2 R : \grave{}(\ ) < 1\}$.

Proof. 1. Since $\grave{}$ is nite everywhere, it is continuous (thus lower semi-continuous), and thus $\grave{} = \grave{}$ and $\grave{}$ is convex lower semi-continuous (Rockafellar, 1970, Theorem 12.2).

2. For any $x 2 R$ and subgradient $g_x 2 @\grave{}(x)$, $\grave{}(0) \quad \grave{}(x) + g(0 \quad x)$. Since $\lim_{z \uparrow 1} \grave{}(z) = 0$ and $\grave{}$ is convex, it follows that $\grave{}$ is nondecreasing, meaning $g \quad 0$, and thus, for any < 0,

$$\grave{}(\ ) = \sup_{x 2 R} \{ x \quad \grave{}(x)\} \quad \sup_{x 2 R} \{ x \quad \grave{}(0) \quad g_x x\} \quad \sup_{x < 0} \{(\quad g_x) x\} \quad \grave{}(0) = 1 :$$

Additionally, since $\inf_x \grave{}(x) = 0$,

$$(0) = \sup_x \{0 \quad x \quad \grave{}(x)\} = \quad \inf_x \grave{}(x) = 0 :$$

3. Next, suppose $\grave{}$ has tightest Lipschitz parameter $2 [0; 1]$, whereby the any subgradient g_x at a point x satis es $|g_x| \quad$. Consequently, proceeding just as in the study of the case < 0, for any $>$ (which is vacuous when $= 1$),

$$\grave{}(\ ) \quad \sup_{x 2 R} \{ x \quad \grave{}(0) \quad g_x x\} \quad \sup_{x > 0} \{(\quad) x\} \quad \grave{}(0) = 1 :$$

On the other hand, let $^0 2 (0;)$ be arbitrary, whereby there must exist $x > y$ with

$$r < \frac{\grave{}(x) \quad \grave{}(y)}{x \quad y}$$

(where the absolute values were dropped since $x > y$ and $\grave{}$ is nondecreasing). Taking any $h 2 @\grave{}(x)$, note

$$r < \frac{\grave{}(x) \quad \grave{}(y)}{x \quad y} \quad \frac{\grave{}(x) \quad (\grave{}(x) + h(y \quad x))}{x \quad y} = h:$$

Consequently, by the Fenchel-young inequality,

$$\grave{} (h) = hx \quad \grave{}(x) < 1 :$$

Since $\grave{}$ is convex, it is nite over a convex set. Sincer was arbitrary, it follows that $\grave{}$ is nite over $[0;)$. If $= 1$, the proof is done; otherwise, when < 1, since $\grave{}$ is lower semi-continuous and nite over $[0;)$, it must also hold that $\grave{} ( ) < 1$.

4. Let $g 2 \grave{}(0)$ be given; by the Fenchel-Young inequality and $\grave{}(0) > 0$,

$$\grave{} (g) = 0 \quad g \quad \grave{}(0) < 0:$$

Since $\grave{} (0) = 0$ and $\grave{}$ is closed and convex, the rst part follows. For the second part, since $\grave{}$ is closed and convex $g 2 @\grave{}(0)$ implies $0 2 @\grave{}(g)$ (Rockafellar, 1970, Theorem 23.5), which is precisely the rst order optimality condition (Borwein and Lewis, 2000, Proposition 3.1.5).

5. Since $\grave{}$ is strictly convex, then $\grave{}$ is continuously di erentiable on $\text{int}(\text{dom}(\grave{}))$ (Hiriart-Urruty and Lemarechal, 2001, Theorem E.4.1.1). Since $\grave{}$ is di erentiable, then $\grave{}$ is strictly convex along every convex subset of $\grave{}^0(R)$ (Hiriart-Urruty and Lemarechal, 2001, Theorem E.4.1.2). Now consider any $2 \text{int}(\text{dom}(\grave{}))$; since $\grave{}$ is lower semi-continuous (as established above), $= r \grave{} (r ( ) = r \grave{}(r \grave{} ( ))$ (Hiriart-Urruty and Lemarechal, 2001, Corollary E.1.3.6, Proposition E.1.4.3), which means $r \grave{}(R) \quad \text{int}(\text{dom}(\grave{}))$. Since $\grave{}$ is convex, $\text{int}(\text{dom}(\grave{}))$ is a convex subset of $r \grave{}(R)$, which means, by the above, that $\grave{}$ is strictly convex on $\text{int}(\text{dom}(\grave{}))$.

   It remains to be shown that $\grave{}$ is strictly convex on $\text{dom}(\grave{})$, note just the interior. As such, let $_1 < _2$ and $2 (0; 1)$ be given with $_1; _2 2 \text{dom}(\grave{})$. Set $_3 := _1 + (1 \quad) _2 2 \text{int}(\text{dom}(\grave{}))$, whereby

$$\frac{_1 + _3}{2} 2 \text{int}(\text{dom}(\grave{})) \quad \text{and} \quad \frac{_2 + _3}{2} 2 \text{int}(\text{dom}(\grave{}));$$

thus

$$\begin{aligned}
\grave{}(\quad_1+(1\quad)_2) &= \grave{}\left(\frac{_1+_3}{2}+(1\quad)\frac{_2+_3}{2}\right)\\
&< \grave{}\left(\frac{_1+_3}{2}\right)+(1\quad)\grave{}\left(\frac{_2+_3}{2}\right)\\
&= \grave{}\left(\frac{(1+\ )_1+(1\quad)_2}{2}\right)+(1\quad)\grave{}\left(\frac{_1+(2\quad)_2}{2}\right)\\
&\quad\frac{(1+\ )}{2}+\frac{(1\quad)}{2}\ \grave{}(_1)\\
&\quad+\ \frac{(1\quad)}{2}+\frac{(1\quad)(2\quad)}{2}\ \grave{}(_2)\\
&= \grave{}(_1)+(1\quad)\grave{}(_2):
\end{aligned}$$

Therefore, to show $\grave{}$ is strictly convex over $\mathrm{dom}(\grave{})$, first choose $_1 = 0$ in the above derivation, whereby $\grave{}$ is strictly convex over $\mathrm{int}(\mathrm{dom}(\grave{})) [f\ 0g$. If $\mathrm{dom}(\grave{}) = R_+$, the proof is done; otherwise, as established above, $\mathrm{dom}(\grave{}) = [0\ ;\ ]$, and the result follows by choosing $_2 = $.

□

Note that loss structure presented in Section 1.4 provides a subset of the guarantees from above.

Proof of Proposition 1.4.2. This result is a combination of various statements in Lemma 1.B.1; in particular, the uniqueness of the optimum $\grave{}^0(0)$ is from strict convexity of $\grave{}$. □

Next, note that L^b also possesses reasonable convexity structure; the relevant part to this chapter is the existence and form of Gâteaux derivatives.

Proposition 1.B.2. Let convex $\grave{} : R \to R_+$ be given, along with empirical measure b over $X \quad f\ 1; +1g$ (corresponding to some finite sample $f(x_i; y_i)g_{i=1}^m$), and corresponding map L^b defined as $L^b(f) = \int \grave{}(f(z))\,db(z)$, and hypothesis class H with corresponding map $A : \to L^1(b)$ be given.

1. L^b and $L^b \circ A$ are convex.
2. L^b is continuous over $L^1(b)$, and $L^b \circ A$ is continuous over .

3. $L^b(p) = \int \ell^*(mp)\,db$ and $L^b(0) = 0$. Furthermore, if ℓ has tightest Lipschitz constant $\lambda := \sup_{x \neq y} |\ell'(x) - \ell'(y)|/|x - y|$, then $L^b(p) < \infty$ if $p \in L^1(b)$ satisfies $p \in [0, 1/m]$ everywhere.

4. Given $f \in L^1(b)$, then $p \in \partial L^b(f)$ if $p \in \partial \ell(f)/m$ everywhere. Additionally, given $A \in \mathcal{A}$, $g \in \partial(L^b \circ A)(\alpha)$ if $g \in A^* \partial L^b(A\alpha)$.

5. If ℓ is differentiable, then for any $f \in L^1(b)$ the Gâteaux derivative $\nabla L^b(f)$ exists (and is an element of $L^\infty(b)$), and moreover $(\nabla L^b(f))_{x_i, y_i} = \ell'(f(x_i, y_i))/m$ everywhere. Similarly, again assuming differentiability of ℓ, then for any $\alpha \in \mathcal{A}$ the Gâteaux derivative $\nabla(L^b \circ A)(\alpha)$ exists (now an element of $L^\infty(\mu)$), and moreover satisfies $\nabla(L^b \circ A)(\alpha) = A^* \nabla L^b(A\alpha)$ and for any $\alpha' \in \mathcal{A}$

$$\begin{aligned} \left\langle \nabla(L^b \circ A)(\alpha), \alpha' \right\rangle &= \frac{1}{m} \sum_{i=1}^m \sum_{h \in H} y_i \ell'((A\alpha)_{x_i, y_i}) h(x_i) \alpha'(h) \\ &= \frac{1}{m} \sum_{h \in H} \sum_{i=1}^m y_i \ell'((A\alpha)_{x_i, y_i}) h(x_i) \alpha'(h). \end{aligned}$$

   Evaluated at a coordinate $e_h \in \mathcal{A}$ (for $h \in H$), this simplifies to

$$\left\langle \nabla(L^b \circ A)(\alpha), e_h \right\rangle = \frac{1}{m} \sum_{i=1}^m y_i \ell'((A\alpha)_{x_i, y_i}) h(x_i).$$

6. If ℓ is differentiable and strictly convex, then L^b is strictly convex, and continuously differentiable along the interior of its domain $\operatorname{dom}(L^b) = \{p \in L^1(b) : L^b(p) < \infty\}$.

Proof. 1. Let $f, f' \in L^1(b)$ and $\lambda \in [0, 1]$ be given; since ℓ is convex,

$$\begin{aligned} \int \ell(\lambda f(z) + (1 - \lambda) f'(z))\,db(z) &\leq \int (\lambda \ell(f(z)) + (1 - \lambda) \ell(f'(z)))\,db(z) \\ &= \lambda \int \ell(f(z))\,db(z) + (1 - \lambda) \int \ell(f'(z))\,db(z). \end{aligned}$$

   For $L^b \circ A$, let $\alpha, \alpha' \in \mathcal{A}$ and $\lambda \in [0, 1]$ be given; since $A : \mathcal{A} \to L^1(b)$ is linear (cf. Lemma 1.A.1),

$$A(\lambda \alpha + (1 - \lambda) \alpha') = \lambda A\alpha + (1 - \lambda) A\alpha',$$

   whereby convexity now follows by convexity of L^b; alternatively, $A\alpha \in L^1(b)$ and $A\alpha' \in L^1(b)$

may be plugged into the above derivation.

2. For L^b, let $f \in L^1(b)$ and > 0 be given; it suffices to exhibit > 0 such that $\|f - f^0\|_1 <$ implies $|L^b(f) - L^b(f^0)| < $. To this end, define real $M := \max_{i \in [m]} |f(x_i, y_i)|$ and compact set $S := [-M - 1, +M + 1]$; since ℓ is convex over $\mathbb{R}$, it has a Lipschitz constant L over S (Hiriart-Urruty and Lemaréchal, 2001, Theorem B.3.1.2). With these definitions in place, choose $:= \min\{1/m, /L\}$. Now let $f^0 \in L^1(b)$ with $\|f - f^0\|_1$ be arbitrary. It follows by $1/m$ and choice of S that f^0, evaluated over the sample, lies within S. Secondly, since $/L$,

$$\begin{aligned} |L^b(f) - L^b(f^0)| &\leq \frac{1}{m} \sum_{i=1}^m |\ell(f(x_i, y_i)) - \ell(f^0(x_i, y_i))| \\ &\leq \frac{L}{m} \sum_{i=1}^m |f(x_i, y_i) - f^0(x_i, y_i)| = L\|f - f^0\|_1 \quad . \end{aligned}$$

It follows that L^b is continuous. This also grants continuity of $L^b \circ A$, since A is a continuous linear operator (cf. lemma 1.A.1).

3. For any $p \in L^1(b)$,

$$\begin{aligned} L^b(p) &= \sup_{f \in L^1(b)} \langle f, p \rangle - L^b(f) \\ &= \frac{1}{m} \sum_{i=1}^m \sup_{f_i \in \mathbb{R}} \left(m p(x_i, y_i) f_i - \ell(f_i) \right) \\ &= \int \ell^*(mp)\, db \end{aligned}$$

(The equivalence classes of functions comprising $L^1(b)$ each contain simple functions; for the general case $L^1(\ )$, more work is necessary to develop the conjugacy relation (cf. Lemma 4.1.2).) From here it directly follows that $L^b(0) = 0$ since Lemma 1.B.1 granted $\ell^*(0) = 0$.

For the second part, if $p(x_i, y_i) \notin [0, /m]$ for some $i \in [m]$, then Lemma 1.B.1 provides $\ell^*(mp(x_i, y_i)) = -1$, whereby the above conjugacy rule provides $L^b(p) = -1$. On the other hand, if $p \in [0, /m]$ everywhere, then Lemma 1.B.1 provides $\ell^*(mp)$ is finite everywhere, and the above conjugacy relation gives $L^b(p) < -1$.

4. Given $f \in L^1(b)$, by the Fenchel-Young inequality, a function $p \in L^1(b)$ satisfies $p \in \partial L^b(f)$ if $L^b(f) + L^b(p) = \langle p, f \rangle$ (Zalinescu, 2002, Theorem 2.4.2.iii). Writing this out, $p \in \partial L^b(f)$ if

$$\sum_i p(x_i, y_i) f(x_i, y_i) = \frac{1}{m} \sum_i \ell(f(x_i, y_i)) + \frac{1}{m} \sum_i \ell^*(mp(x_i, y_i)).$$

For every i with $p(x_i, y_i) \in \partial(\ell(f(x_i, y_i))/m$, Fenchel-Young again grants $p(x_i, y_i) f(x_i, y_i) = \ell(f(x_i, y_i)) + \ell^*(mp(x_i, y_i))$ (Zalinescu, 2002, Theorem 2.4.2.iii). On the other hand, if some i has $p(x_i, y_i) \notin \partial(\ell(f(x_i, y_i))/m$, then Fenchel-Young holds with strict inequality (Zalinescu, 2002, Theorem 2.4.2.iii); consequently $p \in \partial(\ell \circ f)/m$ everywhere.

For $\partial(L^b \circ A)$, since A is a continuous linear operator (cf. Lemma 1.A.1), and since it was established above that L^b is continuous everywhere on $L^1(b)$, the subdifferential rule $\partial(L^b \circ A)(\cdot) = A^* \partial L^b(A \cdot)$ holds (Borwein and Zhu, 2005, Theorem 4.3.3).

5. When ℓ is differentiable, ∂L^b now holds a single equivalence class of b-a.e. equivalent functions, and similarly for $\partial(L^b \circ A)$, which means a Gâteaux derivative exists in both cases, and each time is equal to the corresponding equivalence class identified in the subdifferential rule (Borwein and Zhu, 2005, Theorem 4.2.9).

In order to simplify the expression for $\nabla(L^b \circ A)(\cdot)$ for $\in$, which is an element of $L^1(\ )$, by the earlier dual pairing and the above expansion of L^b, for any $^0 \in$,

$$\begin{aligned}
\left\langle \nabla(L^b \circ A)(\cdot), {}^0 \right\rangle &= \left\langle A^* \nabla L^b(A \cdot), {}^0 \right\rangle \\
&= \left\langle \nabla L^b(A \cdot), A {}^0 \right\rangle \\
&= \sum_{i=1}^m (\ell'((A \cdot)_{x_i, y_i})/m) \left(y_i \sum_{h \in H} h(x_i) {}^0(h) \right) \\
&= \frac{1}{m} \sum_{i=1}^m \sum_{h \in H} y_i \ell'((A \cdot)_{x_i, y_i}) h(x_i) {}^0(h) \\
&= \frac{1}{m} \sum_{h \in H} \sum_{i=1}^m y_i \ell'((A \cdot)_{x_i, y_i}) h(x_i) {}^0(h)
\end{aligned}$$

where Fubini's theorem justifies swapping the final two summations, which is easily valid since $\| {}^0 \|_1 < 1$ and $\sup_{h,x} |h(x)| \leq 1$. Specializing this expression for $\in h \in$ is direct: at

any coordinate e_h, by the above rule for $r\hat{L}b$ and properties of the adjoint $A^>$,

$$\left\langle r(\hat{L}b\ A)(\ );e_h\right\rangle = \left\langle A^> r\hat{L}b(A\ );e_h\right\rangle = \left\langle r\hat{L}b(A\ );h\right\rangle = \frac{1}{m}\sum_{i=1}^{m} y_i\grave{}^0((A\ )_{x_i;y_i})h(x_i):$$

6. By Lemma 1.B.1, $\grave{}$ is strictly convex and continuously di erentiable over the interior of its domain. Thus, given $\ 2 (0;1)$ and $p_1; p_2 2 \mathrm{dom}(\hat{L}b)$ with $p_1 \neq p_2$, there must be at least one example $fx_i; y_i g$ upon which $p_1(x_i; y_i) \neq p_2(x_i; y_i)$, whereby the above conjugacy relation grants

$$\begin{aligned}\hat{L}b(\ p_1 + (1\ )p_2) &= \frac{1}{m}\sum_{i=1}^{m}(\grave{}(\ mp_1(x_i; y_i) + (1\ )mp_2(x_i; y_i)))\\ &< \frac{1}{m}\sum_{i=1}^{m}(\grave{}\ (mp_1(x_i; y_i)) + (1\ )\grave{}\ (mp_2(x_i; y_i)))\\ &= \hat{L}b(p_1) + (1\ )\hat{L}b(p_2):\end{aligned}$$

Next, given some $f 2 L^1(b)$, the same derivation as for $r\hat{L}b$ provides that $r\hat{L}b(f)_{x_i;y_i} = r\grave{}(f(x_i; y_i))=m$, whereby $r\hat{L}b$ is continuously di erentiable since Lemma 1.B.1 granted that $r\grave{}$ is continuously di erentiable.

□

Notice that this appendix makes no assertions regarding $\hat{L}$. While the above results for $\grave{}$ and $\hat{L}b$ may seem remedial, the analogous properties for $\hat{L}$ are not simply more di cult to prove due to technicalities (cf. Lemma 4.1.2), but even basic properties such as niteness of $\hat{L}$ are false in the generality stated above for $\hat{L}b$ (cf. Proposition 4.1.3). Note that another way to obtain fairly general guarantees is to relax the niteness condition on the measure, but instead require H to be nite (cf. Lemma 5.A.2). Consequently, things only break down (for the choice of Banach spaces here) in the case that both the measure over $X\ f\ 1;+1g$ and the measure over H have in nite support.

Lastly, the weak duality result may be proved as follows.

Proof of Proposition 1.4.1. To start, since $\grave{}$ is convex and nite, it is continuous and thus measurable. Moreover, given any $p 2 L^1(\ )$ (whereby $\grave{}\ p$ is measurable), since $\grave{}$ is bounded below by Lemma 1.B.1 and since is a nite measure, $\left|\int_{[\grave{}(p)<0]} \grave{}(p)d\ \right| < 1$. Consequently, only the positive content of $\grave{}\ p$ can lead to an in nite integral, thus $\int \grave{}(p)d$ is well-de ned.

Next, for any $f \in L^1(\ )$ and any $p \in L^1(\ )$, the de nition of Fenchel conjugate provides

$$p(z)f(z) \quad `(f(z)) \quad ` (p(z)):$$

Consequently, for any $p \in L^1(\ )$,

$$Z \qquad Z \qquad Z \qquad Z \qquad Z$$
$$`d \quad (p) = \sup_{f \in L^1(\ )} pfd \qquad `(f)d \qquad \sup_{f \in L^1(\ )} ` (p(z))d\ (z) = \quad ` (p(z))d\ (z);$$

(1.B.3)

where $`$ p is measurable and the corresponding integral is well-de ned as veri ed above.

Now let denote the indicator function over , meaning it is 0 when its argument resides in , otherwise it is in nite. By direct inspection, its conjugate for any $q \in L^1(\ )$ is

$$(q) = \sup_{2} h; q i \qquad (\ ) = \sup_{2} h; q i = \ _{f0g}(q): \tag{1.B.4}$$

For any 2 and $p \in L^1(\ )$, the de nition of Fenchel conjugate provides

$$Z \qquad Z \qquad Z$$
$$` \quad (p) \qquad (A\ )pd \qquad `(A\ );$$
$$Z$$
$$(\ ) \ (A^> p) \qquad (A\ )pd \qquad (A^>\ ):$$

Combining this with eq. (1.B.3) and eq. (1.B.4),

$$Z \qquad Z \qquad Z \qquad Z$$
$$`(A\ ) + \quad (\ ) + \quad ` (p) + \ _{f0g}(\ A^> p) \qquad `(A\ ) + \quad (\ ) + \qquad ` \quad (p) + \quad (\ A^> p)$$
$$Z \qquad Z$$
$$(A\ )p + \quad (\ A^> p)$$

$$= 0:$$

Rearranging,

$$Z \qquad Z$$
$$`(A\ ) + \quad (\ ) \qquad ` (p) \quad \ _{f0g}(\ A^> p):$$

Since 2 and $p \in L^1(\ )$ were arbitrary, $\sup_{p \in L^1(\ )}$ may be applied to the right hand side

and $\inf_{\lambda \geq 0}$ may be applied to the left hand side while preserving the inequality, meaning

$$\inf_{\lambda \geq 0} \left(\int \ell(A\lambda) + \Omega(\lambda) : \lambda \geq 0 \right) \geq \sup_{p} \left(\int \ell^*(p) - \sup_{\{\lambda \geq 0\}} (-A^T p)^T \lambda : p \geq 0, \int |A^T p| \leq 1 \right).$$

In order to obtain the desired statement from here, note first that $\Omega(\lambda) = 0$ for $\lambda \geq 0$, and thus the expression may be dropped. Second, to adjust the dual, first note that $\ell^*(0) = 0$ by Lemma 1.B.1, thus 0 is always a dual feasible point (with value 0), meaning constraints can be added to the dual formulation so long as the value is not forced below zero. To start, note by Lemma 1.B.1 that $\ell^*(r) = \infty$ when $r < 0$, thus the condition $p \geq 0$ μ-a.e. may be safely added, since $\mu([p < 0]) > 0$ entails a dual objective value of $-\infty$. Lastly, the statement $\int_R (Ah) p \, d\mu = 0$ for all h is just a fancy way of writing $\sup_{\{\lambda \geq 0\}} (A^T p)^T \lambda < \infty$. □

Appendix 1.C Line Search Properties

This section provides basic properties of the three line search choices, most importantly a lower bound on the amount they decreased the (primal) objective function in each iteration.

These line searches depend on Lipschitz properties of gradients; concretely, since Boost is coordinate descent applied to $\hat{R} \circ A$, the gradient mapping $\nabla(\hat{R} \circ A)$ will need to be Lipschitz with respect to the $L^1(b)$ norm over a set S, meaning there exists some B_2 so that for any $\lambda, \lambda' \in S$,

$$\| \nabla(\hat{R} \circ A)(\lambda) - \nabla(\hat{R} \circ A)(\lambda') \|_1 \leq B_2 \| \lambda - \lambda' \|_1. \tag{1.C.1}$$

An inequality of this type will be proved separately in Chapter 3 and in Chapter 6, but each time will make use of the following lemma which takes care of the first few steps.

Lemma 1.C.2. Let convex differentiable $\ell : R \to R_+$ be given, along with empirical measure b over $X \times \{-1, +1\}$. Then for any $\lambda, \lambda' \in \Lambda$,

$$\| \nabla(\hat{R} \circ A)(\lambda) - \nabla(\hat{R} \circ A)(\lambda') \|_1 \leq \frac{1}{m} \sum_{i=1}^m |\ell'((A\lambda)_{x_i, y_i}) - \ell'((A\lambda')_{x_i, y_i})| \cdot$$

Proof. By the form of $r(\hat{L}_b A)$ from Proposition 1.B.2 and the form of $\|A^\top\|_{k_1}$ from Lemma 1.A.5,

$$\begin{aligned}
&r(\hat{L}_b A)(\quad) - r(\hat{L}_b A)(\quad) \\
&= \sup\left\{ \frac{1}{m}\sum_{i=1}^m \sum_{h\in H} y_i\left(\ell'((A\quad)_{x_i,y_i}) - \ell'((A\quad)_{x_i,y_i})\right)h(x_i)v(h) \;:\; v\in\quad,\ \|v\|_{k_1}\le 1\right\} \\
&\le \sup\left\{ \frac{1}{m}\sum_{i=1}^m \sum_{h\in H} \left|\ell'((A\quad)_{x_i,y_i}) - \ell'((A\quad)_{x_i,y_i})\right| |y_i h(x_i)|\,|v(h)| \;:\; v\in\quad,\ \|v\|_{k_1}\le 1\right\} \\
&\le \frac{1}{m}\sum_{i=1}^m \left|\ell'((A\quad)_{x_i,y_i}) - \ell'((A\quad)_{x_i,y_i})\right|
\end{aligned}$$

as desired. □

1.C.1 Line Search Options 1 and 2

The following guarantee holds for the unconstrained line search (option 1), and the quadratic upper bound line search (option 2), and moreover provides a derivation of the latter choice.

Lemma 1.C.3. Let hypothesis class H, empirical measure b, integer t denoting an iteration, and convex differentiable loss $\ell : R \to R_+$ be given. Suppose $\hat{L}_b A$ has Lipschitz gradients with parameter B_t with respect to norm $L^1(b)$ over the t^{th} level set

$$S_t := \left\{ \quad : \hat{L}_b(A\quad) \le \hat{L}_b(A_{t-1}) \right\}.$$

(Please see eq(1.C.1) for more on this definition.) Suppose step size η_t is chosen according to one of the first two step choices in Figure 1.3, meaning either $\eta_t = \eta_t$ or $\eta_t \in 2\left[\, r\hat{L}_b(A_{t-1})^\top Av_t / B_t ;\ \eta_t \right]$. Then

$$\begin{aligned}
\eta_t &\ge \frac{\|A^\top\, r\hat{L}_b(A_{t-1})\|_{k_1}}{B_t}; \\
\hat{L}_b(A_t) &\le \hat{L}_b(A_{t-1}) - \frac{\|A^\top\, r\hat{L}_b(A_{t-1})\|_{k_1}^2}{2B_t}.
\end{aligned}$$

Proof. Consider any $\quad \ge 0$ so that $\quad := \quad_{t-1} + v\eta_t \in S_t$. By the form of $r(\hat{L}_b A)$ in Proposition 1.B.2,

$$\int_0^1 \left\langle r\hat{L}_b(A_{t-1} + r(A\quad - A_{t-1})),\ A\quad - A_{t-1} \right\rangle dr = \hat{L}_b(A\quad) - \hat{L}_b(A_{t-1}).$$

As such, by the Lipschitz gradient property (which holds for every element between $_{t\ 1}$ and),

$$\begin{aligned}
\text{Lb}(A(\ _{t\ 1} + v_t)) &= \text{Lb}(A\ _{t\ 1}) + \quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \\
&\quad + \int_0^1 \quad r\ \text{Lb}(A\ _{t\ 1} + r(A \quad A\ _{t\ 1}))\ \ r\ \text{Lb}(A\ _{t\ 1}); A \quad A\ _{t\ 1}\ dr \\
&\quad \text{Lb}(A\ _{t\ 1}) + \quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \\
&\quad + \int_0^1 \quad A^{>} r\ \text{Lb}(A\ _{t\ 1} + r(A \quad A\ _{t\ 1}))\ \ A^{>} r\ \text{Lb}(A\ _{t\ 1})\ _1 \ k v_t k_1 dr \\
&\quad \text{Lb}(A\ _{t\ 1}) + \quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \ + \int_0^1 \quad r\ \ ^2 B_t k v_t k_1^2 dr \\
&\quad \text{Lb}(A\ _{t\ 1}) + \quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \ + \frac{^2 B_t}{2}:
\end{aligned}$$

This nal expression de nes a univariate quadratic with minimum $:= \quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \ =B_t$ (where necessarily $_{t\ 1} + v_t\ 2\ S_t$). This function has slopes everywhere exceeding $-\text{Lb}\ A$ along $[\ _{t\ 1};\ _t]$ (for either choice of step size), and so $\quad _t \quad _t$. (Indeed, these bounds give a derivation for the second step size choices.) To get the second guarantee, note that plugging into the above quadratic and simplifying via

$$\quad A^{>} r\ \text{Lb}(A\ _{t\ 1}); v_t \ ^2 \quad ^2 k A^{>} r\ \text{Lb}(A\ _{t\ 1}) k_1^2$$

gives the desired minimum quadratic upper bound. □

1.C.2 Line Search Option 3: the Wolfe Search

The Wolfe line search is required to output a step size satisfying two conditions; after providing this basic de nition and an analog to the single step guarantee established in Lemma 1.C.3 for the other two line searches, this subsection then closes by providing an e cient means for computing a step size satisfying the Wolfe conditions.

De nition 1.C.4 (Wolfe line search). The Wolfe line search chooses any $_t$ which satis es the

following conditions (where this thesis makes the simple choice $c_1 = 1/3$ and $c_2 = 1/2$):

$$\hat{L}(A(\eta_{t-1} + \eta_t v_t)) \leq \hat{L}(A_{t-1}) + c_1 \left\langle A^\top \nabla \hat{L}(A_{t-1}), v_t \right\rangle \tag{1.C.5}$$

$$\nabla \hat{L}(A(\eta_{t-1} + \eta_t v_t))^\top A v_t \geq c_2 \left\langle A^\top \nabla \hat{L}(A_{t-1}), v_t \right\rangle \tag{1.C.6}$$

This method admits an efficient implementation (cf. Figure 1.4).

The basic guarantee of the Wolfe search is as follows; due to its similarity Lemma 1.C.3, it will be possible to invoke these single-step guarantees, and to then analyze the different step sizes simultaneously.

Lemma 1.C.7. Let hypothesis class $\mathcal{H}$, empirical measure $\hat{b}$, integer t denoting an iteration, and convex differentiable loss $\ell : \mathbb{R} \to \mathbb{R}_+$ be given. Suppose $\hat{L}$ has Lipschitz gradients with parameter B_t with respect to norm $L^1(\hat{b})$ over the t^{th} level set

$$S_t := \left\{ \quad 2 \ : \ \hat{L}(A) \quad \hat{L}(A_{t-1}) \right\}.$$

(Please see eq. (1.C.1) for more on this definition.) Suppose step size η_t satisfies the Wolfe conditions for some $0 < c_1 < c_2 < 1$. Then

$$\eta_t \quad \frac{(1 \quad c_2)\|A^\top \nabla \hat{L}(A^\top{}_{t-1})\|_1}{B_t};$$

$$\hat{L}(A_{t}) \quad \hat{L}(A_{t-1}) \quad \frac{{}^2 c_1 (1 \quad c_2) \|A^\top \nabla \hat{L}(A_{t-1})\|_1^2}{B_t}.$$

Proof. By the first Wolfe condition (cf. eq. (1.C.5)), the step size η_t is a descent step, meaning $\eta_t = \eta_{t-1} + \eta_t v_t \in S_t$, and thus $[\eta_{t-1}, \eta_t] \subseteq S_t$. By the definition of B_t,

$$\left\langle A^\top \nabla \hat{L}(A^\top{}_t) \quad A^\top \nabla \hat{L}(A^\top{}_{t-1}), v_t \right\rangle \leq \left\| A^\top \nabla \hat{L}(A^\top{}_t) \quad A^\top \nabla \hat{L}(A^\top{}_{t-1}) \right\|_1$$

$$\leq B_t \|\eta_t \quad \eta_{t-1}\|_1$$

$$= \eta_t B_t.$$

The rest of the proof is just as for standard Wolfe search guarantees and direct from the Wolfe conditions (Nocedal and Wright, 2006, Theorem 3.2). First, subtracting $\left\langle A^\top \nabla b(A_{t-1}), v_t \right\rangle$ from both sides of eq. (1.C.6) gives

$$\left\langle A^\top \nabla b(A^\top {}_t), A^\top \nabla b(A^\top {}_{t-1}) \right\rangle \geq (c_2 - 1) \left\langle A^\top \nabla b(A^\top {}_{t-1}), v_t \right\rangle,$$

which can be combined with the above derivation to yield

$$\eta_t \geq \frac{(c_2 - 1) \left\langle A^\top \nabla b(A^\top {}_{t-1}), v_t \right\rangle}{B_t} \geq \frac{(1 - c_2) k A^\top \nabla b(A^\top {}_{t-1}) k_1}{B_t}$$

Plugging this into eq. (1.C.5) gives

$$b(A_t) \leq b(A_{t-1}) - \frac{c_1 (1 - c_2) k A^\top \nabla b(A_{t-1}) k_1^2}{B_t}.$$

□

For whatever reason, the Wolfe search is less popular than the backtracking line search, even though the latter requires a guess for the upper bound on the Lipschitz parameter; by contrast, the Wolfe search needs no parameters. Consequently, this thesis elects to explain and investigate the method in a little more generality than is necessary.

Consider any convex differentiable function h over some finite dimensional Euclidean space, a current iterate x, and a descent direction v (that is, $\nabla h(x)^\top v < 0$). By convexity, the linearization of h at x in direction v, symbolically $h(x) + \nabla h(x)^\top v$, will lie below the function. But, by continuity, it must be the case that, for any $c_1 \in (0, 1)$, the ray $h(x) + c_1 \nabla h(x)^\top v$, depicted in Figure 1.5, must lie above h for some small region around x; this gives the first Wolfe condition (presented earlier in eq. (1.C.5)), also known as the Armijo condition (cf. Nocedal and Wright (2006, Equation 3.4) and Bertsekas (1999, Exercise 1.2.16)):

$$h(x + v) \leq h(x) + c_1 \nabla h(x)^\top v. \tag{1.C.8}$$

Unfortunately, this rule may grant only very limited decrease in objective value, since > 0 can be chosen arbitrarily small and still satisfy the rule; thus, the second Wolfe condition (presented earlier in eq. (1.C.6)), also called a curvature condition, which depends on $c_2 \in (c_1, 1)$, forces the

Routine. Wolfe .
Input. Convex function h, iterate x, descent direction v.
Output. Step size > 0 satisfying (1.C.8) and (1.C.9).

1. Bracketing step.
   (a) Set $_{max}$:= 1.
   (b) While $_{max}$ satis es (1.C.8):
       Set $_{max}$:= 2 $_{max}$.
2. Bisection step.
   (a) Set $_{min}$:= 0 and := $_{max}$ =2.
   (b) While does not satisfy (1.C.8) and (1.C.9):
       i. If violates (1.C.8):
          Set $_{max}$:= .
       ii. Else, must violate (1.C.9):
          Set $_{min}$:= .
       iii. Set := ($_{min}$ + $_{max}$)=2.
   (c) Return .

Figure 1.4. Bracketing and bisecting search for step size satisfying Wolfe conditions.

step to be farther away:

$$r\ h(x + \ v)^{>} v \quad c_2 r\ h(x)^{>} v: \tag{1.C.9}$$

This requires the new gradient (in direction v) to be closer to 0, mimicking rst order optimality conditions for the exact line search. Note that the new gradient (in direction v) may in fact be positive; this does not a ect the analysis.

An algorithm to nd a point satisfying these conditions, presented in Figure 1.4, is simple enough: grow as quickly as possible, and then bisect backwards for a satisfactory point. As compared with the presentation in Nocedal and Wright (2006, Algorithm 3.5), $_{max}$ is searched for rather than provided, and convexity removes the need for interpolation.

Proposition 1.C.10. Given a continuously di erentiable convex bounded below function h, iterate x, and direction v, Wolfe terminates with an > 0 satisfying (1.C.8) and (1.C.9).

Proof. The bracketing search must terminate: v is a descent direction, so the linearization at $_{t\ 1}$ with slope $c_1 r\ h(x)^{>} v$ will eventually intersect h (since h it is bounded below).

The remainder of this proof is illustrated in Figure 1.5. Let $_1$ be the greatest positive

Figure 1.5. The mechanism behind Wolfe : the set of points satisfying eq. (1.C.8) and eq. (1.C.9) is a closed interval, and bisection will nd interior points. In this gure, dashed lines denote various relevant slopes.

real satisfying (1.C.8); due to convexity, every 0 satisfying this rst condition must also satisfy 2 [0; 1]. Crucially, 1 < max .

Next, let 2 be the smallest positive real satisfying(1.C.9); existence of such a point follows from the existence of points satisfying both Wolfe conditions (Nocedal and Wright, 2006, Lemma 3.1). By convexity,

$$\langle \nabla h(x + \alpha v) - \nabla h(x), v \rangle \geq 0,$$

and therefore every 0 satisfying (1.C.9) must satisfy 2.

Finally, 1 6= 2, since $c_1 < c_2$, meaning

$$\nabla h(x + \alpha_1 v)^\top v = c_2 \nabla h(x)^\top v < c_1 \nabla h(x)^\top v < \nabla h(x + \alpha_2 v)^\top v.$$

Combining these facts, the interval [2; 1] is precisely the set of points which satisfy (1.C.8) and (1.C.9). The bisection search maintains the invariants min 2 and max 1, meaning no valid solution is ever thrown out: [2; 1] [min ; max]. [2; 1] has nonzero width (since 1 6= 2), and every bisection step halves the width of [min ; max], thus the procedure terminates. □

Appendix 1.D Existence of Hard Cores

In order to prove Theorem 1.7.4, this section first proves the countable additivity of $S_D(H;\ )$, and thereafter proves existence via an optimization on the size of elements in $S_D(H;\ )$ (a similar technique can be used as the first step of the proof of the Hahn Decomposition theorem (Folland, 1999, Theorem 3.3)).

Lemma 1.D.1. Given any $\{c_i\}_{i=1}^{1}$ with $c_i \geq 0$ and $\{p_i\}_{i=1}^{1}$ with $p_i \in D(H;\ )$ and $\sum_i c_i \|p_i\|_1 < 1$, the limit object $p_1 := \sum_i c_i p_i$ exists, and satisfies $p_1 \in D(H;\ )$.

Proof. Let $\{c_i\}_{i=1}^{1}$ and $\{p_i\}_{i=1}^{1}$ be given as specified. First, by the monotone convergence theorem and since $\sum_i c_i \|p_i\|_1 \leq \sum_i c_i \|p_i\|_1 < 1$, the function $p_1 = \sum_i c_i p_i$ exists (i.e., all limits converge pointwise), is measurable, and satisfies $\int p_1 = \sum_i c_i \int p_i < 1$, meaning $p_1 \in L^1(\ )$ (Folland, 1999, Theorem 2.15). Moreover, $\|p_1\|_1 \leq \sum_{i=1}^{1} c_i \|p_i\|_1 < 1$, thus $p_1 \in L^1(\ )$ as well. Now let any $\in R^n$ be given; note that $\sum_i \int |c_i p_i(H\ )| \leq \| \| \|H\| \sum_i \|c_i p_i\|_1 < 1$. Thanks to this, by the dominated convergence theorem (Folland, 1999, Theorem 2.25),

$$\begin{aligned}
\int p_1(x;y)y(H\ )\,d\ (x;y) &= \int \sum_{i=1}^{1} c_i p_i(x;y)y(H\ )\,d\ (x;y) \\
&= \sum_{i=1}^{1} \int c_i p_i(x;y)y(H\ )\,d\ (x;y) \\
&= \sum_{i=1}^{1} c_i \int p_i(x;y)y(H\ )\,d\ (x;y) \\
&= 0.
\end{aligned}$$

□

Lemma 1.D.2. $S_D(H;\ )$ is closed under countable unions.

Proof. Let any collection $\{C_i\}_{i=1}^{1}$ with $C_i \in S_D(H;\ )$ and corresponding weighting $p_i \in D(H;\ )$ be given. Define

$$C := \bigcup_{i=1}^{1} C_i \qquad \text{and} \qquad p := \sum_{i=1}^{1} \frac{p_i}{2^i \max\{1, \|p_i\|_1\}}.$$

By Lemma 1.D.1, p is well-defined and satisfies $p \in D(H;\ )$. Note further that $C = [p > 0]$, and thus $C \in S_D(H;\ )$. □

Proof of Theorem 1.7.4. Consider the optimization problem

$$d := \sup\{ (C) : C \in S_D(H;)\}:$$

Since S_D is nonempty (always contains; corresponding to $p = 0 \in D(H;))$ and $(X \{ 1; +1\}) < 1$, the supremum is finite. Let $\{C_i\}_{i=1}^{1}$ be a maximizing sequence, and define $D_j := [_{i \ j} C_i$ and $D := [_{j=1}^{1} D_j = [_{i=1}^{1} C_i$. By Lemma 1.D.2, $D_j \in S_D(H;)$ for every j, and since $(D_j) \ (C_j)$, it follows that $\{D_j\}_{j=1}^{1}$ must also be a maximizing sequence to the above supremum. Finally, since Lemma 1.D.2 also grants $D \in S_D(H;)$, then by continuity of measures from below (Folland, 1999, Theorem 1.8(c)),

$$(D) = \lim_{j \to 1} (D_j) = d:$$

Since $D \in S_D(H;)$ attains the supremum, it is a hard core; in particular, hard cores exist.

If there are two hard cores D_1 and D_2, then the definition provides that each satisfies $D_i \in S_D(H;)$, and $(D_2 \setminus D_1) \ 0$ and $(D_1 \setminus D_2) \ 0$, meaning the symmetric difference has measure zero, which gives the second property. □

Appendix 1.E Bibliographic Notes

As discussed in Section 1.1, the original boosting algorithm is due to Schapire (1990), who also gave the basic guarantees closing the question posed by Kearns and Vazirani (1994). The complexity of the circuit produce by this algorithm was reduced to a single majority by Freund (1995), and thereafter both authors together developed AdaBoost, which is the topic of study in this thesis (Freund and Schapire, 1997). Most recently, Schapire and Freund have also put out an excellent, comprehensive book on many aspects of AdaBoost (Schapire and Freund, 2012). It is noteworthy that while boosting variants abound, practical boosting implementations are still very close to the original (Caruana and Niculescu-Mizil, 2006).

The algorithm was not originally phrased as minimization of a convex loss; this view appears to have been first presented by (Breiman, 1999). Subsequently, the algorithm was studied for many alternate losses, primarily the logistic loss, by many authors Friedman et al. (2000), Friedman (2000), Lafferty (1999), Mason et al. (2000), Collins et al. (2002), Duffy and Helmbold (2000).

Occasionally, authors express the view that the loss minimization view of boosting is somehow incomplete or inaccurate. One reason for this is that the primal optimization problem given in Equation (1.3.4) does not place any restrictions on the complexity of the learned function; since H is typically chosen to be some large VC class, its linear span $\operatorname{span}(H)$ typically has in nite VC dimension, which means it is statistically unstable (Devroye et al., 1996, Theorem 14.3). Consequently, if the objective function is not modi ed in some way, statistical stability must be achieved via some aspect of the minimization algorithm. Boosting in turn accomplishes this in two ways: rst via coordinate descent, which means the t^{th} iterated is supported on at most t coordinates, and secondly via some sort of stopping rule.

Another reason why focusing on the primal objective function is perhaps distasteful is that it omits any mention of the weights over examples and the basic computational model of boosting. This thesis aims to point out that the convex dual highlights that the weighting structure was present in the convex formulation all along.

It was arguably a great boon that the original authors did not formulate the algorithm with loss minimization in mind. Had that been the case, it would have been more expedient to analyze the scheme under the settings more common for convex optimization (strong convexity, existence of minima, etc). As discussed throughout this chapter, and highlighted in the problem structure in Chapter 2, the typical convex case, and the original case of boosting (> 0), are very di erent. As such, thanks to the original algorithm not being phrased as loss minimization, the world now has these two very di erent analyses, which can be understood via hard cores to rather fundamentally partition the space of possibilities.

Regarding hard cores, the term itself is from cryptography (Goldreich and Levin, 1989). Closer to the spirit here, Impagliazzo (1995) showed via a \hard core lemma" that the existence of functions which are somewhat hard for circuits of a given size implies the existence of functions which are very hard for circuits of a smaller size. Similarly, the hard core C of an instance here is a subset of the inputs which is very di cult, and captures the basic di culty of the problem (in Chapter 3 and Chapter 6, when the hard core is empty, the instance is easy). Interestingly, the only existing proofs of fast numerical convergence of AdaBoost to $L^b(A)$ use hard cores; one example is Chapter 3, and the other is the proof due to Mukherjee et al. (2011, See \zero loss set", the complement of the hard core). The hard core may also be used to develop a theory of margins when separability does not hold (Telgarsky, 2013).

Acknowledgements

This chapter is based on work by the dissertation author which appeared in the Journal of Machine Learning Research, Volume 13, pages 561{606, 2012, as well as the Conference on Learning Theory, 2013, and lastly work under preparation.

Part I

Just a Finite Sample

Chapter 2

Problem Structure

This chapter develops the structure of the primal optimization problem over finite samples (e.g., eq. (1.3.3) with $\mathbb{P} = \mathbb{P}_b$), which is then used in Chapter 3 to produce numerical convergence rates. Concretely, these two chapters operate under the following two assumptions.

1. Only the empirical measure $\mathbb{P}_b$ corresponding to a finite sample $\{(x_i, y_i)\}_{i=1}^m$ is considered. The material in Chapter 5 may be used produce statistical guarantees (i.e., to infer behavior over $\mathbb{P}$ from the behavior over $\mathbb{P}_b$).

2. H must satisfy the following notion of bounded complexity: the set of images $\{(h(x_i))_{i=1}^m : h \in H\}$ has some finite size $n < \infty$. This condition is satisfied if for instance the set of possible images $\{h(x) : x \in X, h \in H\}$ has some finite size $k < \infty$ (whereby the above image set has size at most k^m), or for instance when H has finite VC dimension $V(H) < \infty$ (whereby the Sauer-Shelah lemma grants that the size is eventually $O(m^{V(H)})$).

These two assumptions together entail that the linear operators H and A may be represented as matrices in $\mathbb{R}^{m \times n}$ without affecting any of the mathematical development; indeed, to succinctly state these two assumptions, the results will require $A \in \mathbb{R}^{m \times n}$. In this setting, $L^1(\mathbb{P}_b)$ is isometrically isomorphic to $\mathbb{R}^m$ with the l_1 norm; thus for any vector $\alpha \in \mathbb{R}^n$, the object $A\alpha \in \mathbb{R}^m$ is simultaneously considered as an element of $\mathbb{R}^m$, and also an element of $L^1(\mathbb{P}_b)$ via isometric isomorphism. Similarly, note the simplified notation $(A\alpha)_i := (A\alpha)_{x_i, y_i}$.

2.1 Overview

This chapter will study the basic structure of the optimization problem for $L^1_{\mathbb{P}_b}$ under the condition $A \in \mathbb{R}^{m \times n}$. All results will hold for the following family of loss functions.

De nition 2.1.1. Let L_{fs} contain all functions $` : R ! R_{++}$ which are strictly convex, continuously di erentiable, and satisfy $\lim_{z! 1} `(z) = 0$.

The rst step, in Section 2.2, is to provide a strong duality analog to Proposition 1.4.1, but now under the assumption of a nite sample. This result will be proved under slightly more generality than stated above; in particular, the measure will be b, but may be in nite. Even so, focusing on b means that only L^b and L^b must be considered, which were shown to be well behaved in Proposition 1.B.2. Moreover, the dual space, rather than being $L^1( )$ (which is a bad choice for the exponential loss, as discussed in Chapter 4), is simply R^m (where all metric topologies agree since the space has nite dimension).

Section 2.2 will also esh out the duality-based stopping condition in Figure 1.3, which is simpli ed via a few more de nitions. First note the following kernel and image de nitions for A and $A^>$, which are stated in full generality, but will only be used for $A 2 R^{m n}$.

De nition 2.1.2. Given a linear operator $B : X ! R$, de ne its image and kernel as

$$\operatorname{im}(B) := f Bx : x 2 X g \quad \text{and} \quad \ker(B) := f x 2 X : Bx = 0 g:$$

Notice for instance that signed decorrelating weightings are simply $\ker(A^>)$; the following de nition in turn ensures nonnegativity and simpli es many future expressions.

De nition 2.1.3. Given operator $A : L^1( ) ! R$, and identifying $\ker(A)$ with a subset of $L^1( )$ via isometric isomorphism as per Lemma 1.A.5), de ne $D_A := \ker(A) \setminus f p 2 L^1( ) : p 0 \text{ -a.e.} g$.

Notice that this expression contains most of the domain of the dual in Proposition 1.4.1; when $A 2 R^{m n}$, the earlier tacit isometric isomorphisms between $L^1(b)$ and R^m allow D_A to be identi ed with $\ker(A^>) \setminus R^m_+$. Note however that L^b is not nite everywhere over D_A; when $` 2 L_{fs}$ is Lipschitz with parameter , then L^b is only nite over the cube $[0; =m]^m$ (cf. Proposition 1.B.2). On the other hand, given $p 2 D_A \setminus f0g$, the rescaling $p := p (=(m kpk_1))$ always satis es $p 2 \ker(A^>)$ and $L^b(p) < 1$.

The following notation for projections and distances will also be used in the duality-based stopping condition.

De nition 2.1.4. Let $C \subseteq R^m$ be a closed convex set, and de ne the following projection and distance shorthands, where $p \in [1; 1]$, and projections in the case of non-uniqueness (l_1 and l_1 norms) make some arbitrary consistent choice:

$$P_C^p(x) \in \arg\min_{y \in C} \|y - x\|_p; \qquad \text{and} \qquad D_C^p(x) = \|x - P_C^p(x)\|_p:$$

Excluding appendices, the chapter has two more sections. First, Section 2.4 pins down the structure of hard cores under the assumption $A \in R^{m \times n}$. Second, Section 2.3 provides a generalization of the quantity (b) which is positive under conditions which will hold for the optimization problem in general, and thus the modi ed quantity will appear in all convergence rates of Chapter 3, and not simply those corresponding to the case $(b) > 0$.

2.2 Strong Duality

Theorem 2.2.1. Let empirical probability measure b, hypothesis class H (nite or in nite), and loss $\ell : R \to R_+$ be given with $\lim_{z \to 1} \ell(z) = 0$. Then

$$\inf \int \ell(A \ )db : \ \in \qquad = \max \int \ell^*(mp)db : \ \in D_A \ ; \tag{2.2.2}$$

and a pair $(\ \in \ ; p) \in \qquad L^1(b)$ is optimal i $p \in D_A$ and $p \in \partial(A \ )=m$ everywhere. Furthermore, if ℓ is di erentiable and strictly convex, then then the dual optimum $\ell_A \in D_A$ is unique, and $\in$ is optimal i $\ell_A = \ell'(A \ )=m$ everywhere.

Since the structure of L^b and L^b was already handled by Proposition 1.B.2, the proof is little more than an invocation of an appropriate Fenchel duality rule and some remedial manipulations. Additionally, recall, as discussed in Remark 1.A.4, that the appearance of m in the dual (which was absent in the weak duality formula from Proposition 1.4.1) is an artifact of using inconsistent dual pairings for L^b and L.

Special notation is used for the dual optimum ℓ_A in this chapter and Chapter 3, since $A \in R^{m \times n}$ will be shown to entail very strict properties. As such, x the notation as follows.

De nition 2.2.3. When $A \in R^{m \times n}$ and $\ell \in L_{fs}$, let ℓ_A denote the dual optimum as in Theorem 2.2.1.

Proof of Theorem 2.2.1. Consider the two Fenchel problems

$$V_p := \inf \int \ell(A\,\cdot)\,db + \quad (\ );$$
$$V_d := \sup_p \int \ell^*(-mp)\,db \quad (p);$$

where $_S$ denotes a convex indicator function (0 on S, $+1$ on S^c), and the conjugacy relation $(\int \ell\,db)^*(p) = \int \ell^*(pm)\,db$ was provided by Proposition 1.B.2.

In order to show $V_p = V_d$, since the dual space $L^1(b)$ is nite dimensional, it su ces to apply a variant of Slater's condition making use of quasi relative interiors (Zalinescu, 2002, Corollary 2.8.5 using condition (x)); namely, the relative interior of $\mathrm{dom}(L^b)$ is simply $L^1(b)$ as provided by Proposition 1.B.2 and the quasi relative interior of includes the origin, so the constraint quali cation is satis ed, whereby $V_p = V_d$, there is attainment in the dual, and 2 is optimal i there exists a dual optimum p with $A^\top p \in \partial(\ )(0) = \{0\}$ and $p \in \partial L^b(A\,\cdot)$. By the form of ∂L^b given in Proposition 1.B.2, $p \in \partial L^b(A\,\cdot)$ i $p \in \partial \ell(A\,\cdot)/m$, and the rst condition can be simpli ed to say $\int \ell(A\,\cdot)p\,db = 0$ for all , which in turn is equivalent to $p \in \ker(A^\top)$. Since $\partial\ell \subseteq \mathbb{R}_+$, it follows that $p \in D_A$. Furthermore, if ℓ is strictly convex and di erentiable, by Proposition 1.B.2 it follows that L^b is also strictly convex whereby p must be unique, and moreover the simpli ed optimality condition follows by the derivative rules given in Proposition 1.B.2.

In order to translate the dual expression into the desired form, rst note that

$$\ell^*(q) = \sup_{\in\mathbb{R}} \{\langle\, ; q\,\rangle - \ell(\ )\} = \ _{\{0\}}(q).$$

As such, since $L^b(0) = 0$ by Proposition 1.B.2 and $\ell(A^\top 0) = 0$ by the above, the dual value is always at least 0, thus values below this may be safely removed from the dual domain. First, $L^b(p) = \infty$ when p is not nonnegative everywhere by proposition 1.B.2, thus $p \geq 0$ may be stipulated. This combined with $_{\{0\}}(A^\top p) = 0$ i $\int \ell(A\,\cdot)p\,db = 0$ for all p means the dual domain may be replaced with D_A.

Lastly, the simpli ed form of the optimality criteria are from Proposition 1.B.2, where the dual optimum is unique by strict convexity of L^b again by Proposition 1.B.2. □

Remark 2.2.4. Continuing the connection to maximum entropy, for any $\ell \in L_{fs}$, by Lemma 1.B.1, $p \mapsto \int \ell^*(-mp)\,db$ attains its maximal value at the vector $p \in \mathbb{R}^m$ with $p_i = \ell'(0)/m$ for all

coordinates $i \in [m]$, which is a rescaling of the uniform distribution. But recalling that the algorithm is initialized with $\lambda_0 = 0$, then Proposition 1.B.2 grants $\nabla \hat{L}(A\lambda_0) = \nabla \hat{L}(0) = p$: that is, the initial dual iterate $p = \nabla \hat{L}(A\lambda_0)$ is the unconstrained optimum! Let $\theta_t := \nabla \hat{L}(A\lambda_t)$ denote the t^{th} dual iterate; since $\nabla \hat{L}^* (\nabla \hat{L}(z)) = z$ for any z (Zalinescu, 2002, Theorem 2.4.2.ii), then for any $\lambda \in D_A \cap \ker(A^\top)$,

$$\left\langle \nabla \hat{L}^*(\theta_t), \lambda \right\rangle = \langle A\lambda_t, \lambda \rangle = \langle \lambda_t, A^\top \lambda \rangle = 0.$$

This allows the dual optimum to be rewritten as

$$\begin{aligned} \hat{\lambda}_A &= \operatorname*{arg\,min}_{\lambda \in D_A} \int \ell(-z) \, db(z) \\ &= \operatorname*{arg\,min}_{\lambda \in D_A} \int \ell(-z) \, db(z) - \int \ell(-z) \, db(z) - \left\langle \nabla \hat{L}^*(\theta_t), \lambda \right\rangle \\ &= \operatorname*{arg\,min}_{\lambda \in D_A} D_{\hat{L}}(\lambda, \lambda_t), \end{aligned}$$

that is, the dual optimum $\hat{\lambda}_A$ is the Bregman projection (according to $\hat{L}$) onto D_A of any dual iterate $\theta_t = \nabla \hat{L}(A\lambda_t)$. In particular, $\hat{\lambda}_A$ is the Bregman projection onto the feasible set of the unconstrained optimum $\theta_0 = \nabla \hat{L}(A\lambda_0)$!

The connection to Bregman divergences runs deep; in fact, just as the specification of Boost in Figure 1.3 is is purely in terms of primal variables (whereas the original formulation used both primal and dual variables), it is also possible to specify the algorithm purely in terms of dual variables, where the line search performs Bregman projections (Kivinen and Warmuth, 1999, Collins et al., 2002).

Before adjourning this section, the above strong duality rule and projection notation together lead to the duality gap stopping criterion (option 2 in Figure 1.3).

Remark 2.2.5. Given any λ_t, the corresponding dual variable $\theta_t := \nabla \hat{L}(A\lambda_t)$ is not necessarily dual feasible, which can be achieved by instead working with the projection $\theta_t := P_S^1(\theta_t)$ where $S := D_A \cap \{p \in \mathbb{R}^m : \hat{L}(p) < 1\}$, the latter being a closed convex set. In order to ensure accuracy $\varepsilon > 0$, it suffices by Theorem 2.2.1 to check $\hat{L}(A\lambda_t) + \hat{L}^*(\theta_t) \leq \varepsilon$.

Unfortunately, this duality gap computation has two significant weaknesses. First, it is painfully expensive to compute; said another way, this projection does not fit the computational model of boosting, where A is never intended to be explicitly constructed or stored. (What is $\ker(A^\top)$ when $\mathcal{H}$ consists of size-bounded decision trees?) Concretely, the primal-dual relationship

in Theorem 2.2.1 can be written

$$\inf\left\{ L^b(\lambda) : \quad \lambda \in \operatorname{im}(A) \right\} = \max\left\{ L^b(-A^\top \mu) : \quad \mu \in \ker(A^\top) = \operatorname{im}(A)^\perp \right\}$$

(since $\operatorname{dom}(L^b) \subseteq \mathbb{R}^m_+$ encodes the orthant constraint), where the computational black box gives elements of $\operatorname{im}(A)$, but what is needed in the dual is a black box for $\ker(A^\top) = \operatorname{im}(A)^\perp$.

The second weakness is that the optimization guarantees of Chapter 3 do not indicate how fast $L^b(A\lambda_t) + L^b(-A\lambda_t)$ decreases. The guarantees themselves only indicate how fast the error $L^b(A\lambda_t) - L^b(A\lambda)$ decreases, and while the proofs do go through a duality gap obtained via projections, these projections are onto a set which does not seem easily computed by the algorithm.

2.3 Generalized Weak Learning Rate

The weak learning rate $\gamma(b)$ was critical to the original convergence analysis of AdaBoost, providing a handle on the progress of the algorithm. But to be useful, this value must be positive, which was precisely the condition granted by the (empirical) weak learning assumption. This section will generalize the weak learning rate $\gamma(b)$ into a quantity which can be made positive for any boosting instance.

Now let $\gamma(b)$ denote the largest possible value for $\gamma(b)$; this corresponds to the optimization problem

$$\gamma(b) = \inf_{\substack{\lambda \in \mathbb{R}^m_+ \\ \|\lambda\|_1 = 1}} \max_{j \in [n]} \sum_{i=1}^m (\lambda_i y_i h_j(x_i)) = \inf_{\lambda \in \mathbb{R}^m_+ \setminus \{0\}} \frac{\|A^\top \lambda\|_1}{\|\lambda\|_1} = \inf_{\lambda \in \mathbb{R}^m_+ \setminus \{0\}} \frac{\|A^\top \lambda\|_1}{\|\lambda\|_1}; \qquad (2.3.1)$$

where the final form, with extraneous zero terms, will lead the path to the eventual generalization. For now, this simplified form leads to the following alternate characterization.

Proposition 2.3.2. A boosting instance is empirically weakly learnable (i.e., it is possible to choose $\gamma(b) > 0$) iff $D_A = \{0\}$.

Proof. Suppose $D_A = \{0\}$; since the first minimum in eq. (2.3.1) is of a continuous function over a compact set, it has some minimizer λ^0. But $\|\lambda^0\|_1 = 1$, meaning $\lambda^0 \notin D_A$, and so $\|A^\top \lambda^0\|_1 > 0$,

and (b) may be chosen positive. On the other hand, if $D_A = \{0\}$, take any $\delta'' \in D_A \setminus \{0\}$; then

$$0 \quad (b) = \inf_{\lambda \in \mathbb{R}^m_+ \setminus \{0\}} \frac{\|A^\top \lambda\|_1}{\|\lambda\|_1} \quad \frac{\|A^\top \delta''\|_1}{\|\delta''\|_1} = 0 :$$

□

As such, to weaken the assumption $(b) > 0$ is the same as moving away from $D_A = \{0\}$, and a way to brute force this is to take various zeroes in the final expression of eq. (2.3.1), and replace them with something else. This generalization depends on a generic set S, which was alluded to in the duality gap discussions in Remark 2.2.5, and will only be set in Chapter 3.

Definition 2.3.3. Given a matrix $A \in \mathbb{R}^{m \times n}$ and a set $S \subseteq \mathbb{R}^m$, define

$$(A; S) := \inf \left\{ \frac{\|A^\top \lambda\|_1}{\inf_{\delta \in S \setminus \ker(A^\top)} \|\lambda\|_1} : \lambda \in S \setminus \ker(A^\top) \right\}$$

First note that in the scenario of weak learnability (i.e., $D_A = \{0_m\}$ by Proposition 2.3.2), the choice $S = \mathbb{R}^m_+$ allows the new notion to exactly cover the old one: $(A; \mathbb{R}^m_+) = (b) \quad (b)$.

To get a better handle on the meaning of S, recall the projection and distance notation P and D from Section 2.1. Suppose, for some t, that $r(b)(A^\top) \in S \setminus \ker(A^\top)$; then the infimum within $(A; S)$ may be instantiated with $r(b)(A^\top)$, yielding

$$(A; S) = \inf_{\lambda \in S \setminus \ker(A^\top)} \frac{\|A^\top \lambda\|_1}{\|\lambda\| \quad P^1_{S \setminus \ker(A^\top)}(\lambda)\|_1} \quad \frac{\|A^\top r(b)(A^\top)\|_1}{\|r(b)(A^\top) \quad P^1_{S \setminus \ker(A^\top)}(r(b)(A^\top))\|_1} : \tag{2.3.4}$$

Rearranging this,

$$(A; S) \quad r(b)(A^\top) \quad P^1_{S \setminus \ker(A^\top)}(r(b)(A^\top)) \quad \|A^\top r(b)(A^\top)\|_1 : \tag{2.3.5}$$

This is helpful because the right hand side appears in standard guarantees for single-step progress in descent methods. Meanwhile, the left hand side has reduced the influence of A to a single number, and the normed expression is the distance to a restriction of dual feasible set, which will converge to zero if the infimum is to be approached, so long as this restriction contains the dual optimum.

This will be exactly the approach taken in Chapter 3; indeed, the first step towards convergence rates, Proposition 3.1.2, will use exactly the upper bound in eq. (2.3.5). The detailed work that remains is then dealing with the distance to the dual feasible set. The choice of S will be made to facilitate the production of these bounds, and will depend on the optimization structure revealed in Section 2.4.

In order for these expressions to mean anything, $(A; S)$ must be positive.

Theorem 2.3.6. Let matrix $A \in \mathbb{R}^{m \times n}$ and polyhedron $S \subseteq \mathbb{R}^m$ be given with $S \cap \ker(A^\top) \neq \emptyset$ and $S \setminus \ker(A^\top) \neq \emptyset$. Then $(A; S) > 0$.

The proof, material on other generalizations of (b), and discussion on the polyhedrality of S can all be found in Section 2.A.

As a final connection, since $A^\top P^1_{S \setminus \ker(A^\top)}(\cdot) = 0$ for any $\cdot$, note that

$$(A; S) = \inf_{\in S \cap \ker(A^\top)} \frac{\|A^\top \cdot\|_1}{\|\cdot - P^1_{S \setminus \ker(A^\top)}(\cdot)\|_1} = \inf_{\in S \cap \ker(A^\top)} \frac{\|A^\top (\cdot - P^1_{S \setminus \ker(A^\top)}(\cdot))\|_1}{\|\cdot - P^1_{S \setminus \ker(A^\top)}(\cdot)\|_1};$$

which now makes use of both extraneous zeros in eq. (2.3.1). To connect this to standard notions from linear algebra, if A were square and full rank, then the relationship between $\|A^\top \cdot\|_2$ and $\|\cdot\|_2$ could be bounded by the smallest and largest singular values of A. The above expression has two modifications: it is a more general operator norm (with l_1 over the domain and l_1 over the image), and explicitly avoids problems caused by A not being full rank.

2.4 The Hard Core

When $A \in \mathbb{R}^{m \times n}$, the hard core may be specified in great detail. By Theorem 1.7.4, hard cores always exist, and since $L^{-1}(b)$ is identified here with $\mathbb{R}^m$, a hard core C is not only unique, but moreover it can be identified simply as a subset of the index set $[m]$.

It will be useful to view the rows $\{a_i\}_{i=1}^m$ of A as a collection of m points in $\mathbb{R}^n$. Due to the form of L^b for $\ell \in L_{fs}$, Boost is therefore searching for a halfspace, parameterized by a vector $$, which contains all of these points. Sometimes such a halfspace may not exist, and applies a smoothly increasing penalty to points that are farther and farther outside it. It will be shown that the best choices of place the entire hard core on the boundary of the halfspace defined by (where the degenerate halfspace corresponding to $= 0$ is understood to have all of $\mathbb{R}^m$ as its boundary), whereas the complement of the hard core falls interior to the halfspace.

2.4.1 Weak Learnability

Recall Proposition 2.3.2, which provided that an instance is empirically weakly learnable i $D_A = f0g$. With this in mind, the following theorem establishes four equivalent formulations of empirical weak learnability.

Theorem 2.4.1. For any $A 2 R^{m n}$, the following conditions are equivalent.

1. $9 2 R^n$ $A 2 R^m$,
2. $8` 2 L_{fs}$, $L^{b}(A) = 0$,
3. $8` 2 L_{fs}$, $\grave{A} = 0$,
4. $D_A = f0g$.

First note that item 4 indicates (via Proposition 2.3.2) that this is indeed captures the setting of weak learnability, which is equivalent to $jCj = 0$ where C is the hard core.

Recall the earlier discussion of boosting as searching for a halfspace containing the points $f a_i g_{i=1}^m = f e_i^> Ag_{i=1}^m$; item 1 encodes precisely this statement, and moreover that there exists such a halfspace with these points interior to it. Note that this statement also encodes the margin separability equivalence of weak learnability due to Shalev-Shwartz and Singer (2008); speci cally, if labels are bounded away from 0 and each point a_i (row of A) is replaced with $y_i a_i$, the de nition of A grants that positive examples will land on one side of the hyperplane, and negative examples on the other.

Item 4 and item 1 can be interpreted geometrically, as depicted in Figure 2.1: the dual feasibility statement is that no convex combination of $f a_i g_{i=1}^m$ will contain the origin.

From this, the fact $L^{b}(A) = 0$ (Item 2) is immediate. Lastly, when $D_A = f0g$, the dual optimum satis es $\grave{A} = 0$ (Item 3).

Proof of Theorem 2.4.1. (Item 1 =) item 2.) Let $2 R^n$ with $A 2 R^m$, any increasing sequence $f c_i g_1^1 " 1$, and any $` 2 L_{fs}$ be given. Then, since $L^{b} > 0$ and $\lim_{x ! 1} `(x) = 0$,

$$\inf L^{b}(A) \lim_{i !1} L^{b}(c_i A) = 0 \inf L^{b}(A):$$

(Item 2 =) item 3.) Let any $` 2 L_{fs}$ be given. The point 0 is always dual feasible, and

Figure 2.1. Geometric view of the primal and dual problem, under weak learnability. The vertices of the pentagon denote the points $\{a_i\}_{i=1}^m$. The arrow, denoting θ in item 1 of Theorem 2.4.1, defines a homogeneous halfspace containing these points; on the other hand, their convex hull does not contain the origin. Please see Theorem 2.4.1 and its discussion.

$\ell(0) = 0$ (cf. Lemma 1.B.1) and the form of L^b (cf. Proposition 1.B.2) grants

$$\inf L^b(A\theta) = 0 = L^b(0).$$

Since the dual optimum is unique (cf. Theorem 2.2.1), $\lambda_A = 0$.

(Item 3 $\Rightarrow$ item 4.) Choose any $\ell \in L_{fs}$, whereby the assumption grants a unique optimum $\lambda_A = 0$. Suppose contradictorily that there exists another dual feasible point $\theta \in \mathbb{R}^m_+$. By Lemma 1.B.1, L^b is negative along $[0, \ell'(0)]^m \cap \{0\}$, which has nonzero intersection with $[\lambda_A, \theta]$; consequently, there must exist $\theta' \in D_A \cap \{0\}$ with $L^b(\theta') > 0 = L^b(\lambda_A)$, which contradicts the optimality of λ_A.

(Item 4 $\Rightarrow$ item 1.) This case is a one direction of Gordan's theorem (Borwein and Lewis, 2000, Theorem 2.2.6). □

2.4.2 Attainability

For strictly convex functions, there is a nice characterization of attainability, which will require the following definition.

Definition 2.4.2 (Hiriart-Urruty and Lemaréchal (2001, Section B.3.2)). A closed convex function h is called *0-coercive* when all level sets are compact. (That is, for any $\alpha \in \mathbb{R}$, the set $\{x : f(x) \leq \alpha\}$ is compact.)

Proposition 2.4.3. Suppose h is strictly convex, closed, and proper. Then $\inf_x h(x)$ is attainable

i h is 0-coercive.

Armed with this notion, it is now possible to build an attainability theory for L^b A. Some care must be taken with the above concepts, however; note that while L^b is strictly convex, L^b A need not be (for instance, if there exist nonzero elements of ker(A), then moving along these directions does not change the objective value). Therefore, 0-coercivity statements will refer to the function

$$(\hat{L}^b + \delta_{\mathrm{im}(A)})(x) = \begin{cases} \hat{L}^b(x) & \text{when } x \in \mathrm{im}(A); \\ \infty & \text{otherwise.} \end{cases}$$

This function is effectively taking the epigraph of L^b, and intersecting it with a slice representing im(A) = f A : 2 R^n g, the set of points considered by the algorithm. As such, it is merely a convenient way of dealing with ker(A) as discussed above.

Theorem 2.4.4. For any A 2 R^{m n}, the following conditions are equivalent.

1. 8 2 R^n A 62 R^m nf0g,
2. 8` 2 L_fs, L^b + im(A) is 0-coercive,
3. 8` 2 L_fs, ` A 2 R^m_{++},
4. D_A \ R^m_{++} 6= ;.

Following the discussion above, item 2 is the desired attainability statement.

Next, note that item 4 is equivalent to the expression jCj = m (where C is the hard core), i.e. there exists a reweighting p of b with p > 0 which decorrelates every h 2 H from the target.

For a geometric interpretation, consider item 1 and item 4. Item 1 says that any halfspace containing some a_i within its interior must also fail to contain some a_j (with i 6= j). (Item 1 also allows for the scenario that no valid enclosing halfspace exists, i.e. = 0.) Item 4 states that the origin 0 is contained within a positive convex combination of f a_i g^m_{i=1} (alternatively, the origin is within the relative interior of these points). These two scenarios appear in Figure 2.2.

Finally, note item 3: it is not only the case that there are dual feasible points fully interior to R^m_+, but furthermore the dual optimum is also interior. This will be crucial in the convergence rate analysis, since it will allow the dual iterates to never be too small.

Figure 2.2. Geometric view of the primal and dual problem, under attainability. Once again, the $\{a_i\}_{i=1}^m$ are the vertices of the pentagon. This time, no (closed) homogeneous halfspace containing all the points will contain one strictly, and the relative interior of the pentagon contains the origin. Please see Theorem 2.4.4 and its discussion.

Proof of Theorem 2.4.4. (Item 1 $\Rightarrow$ item 2.) Let $d \in \mathbb{R}^m \setminus \{0\}$, $\in \mathbb{R}^n$, and $\ell \in L_{fs}$ be arbitrary. To show 0-coercivity, it suffices (Hiriart-Urruty and Lemaréchal, 2001, Proposition B.3.2.4.iii) to show

$$\lim_{t \to \infty} \frac{\bar{L}(A + td) + \sigma_{\mathrm{im}(A)}(A + td) - \bar{L}(A)}{t} > 0. \tag{2.4.5}$$

If $d \in \mathrm{im}(A)$ and $t > 0$, then $\sigma_{\mathrm{im}(A)}(A + td) = 1$. Now suppose $d \notin \mathrm{im}(A)$; by item 1, since $d \neq 0$, then $d \notin \mathbb{R}^m$, meaning there is at least one positive coordinate e_j. But then, since $\ell > 0$ and ℓ is convex,

$$\begin{aligned}(2.4.5) \quad &\lim_{t \to \infty} \frac{m^{-1}\ell(e_j^T(A + td)) - \bar{L}(A)}{t} \\ &\lim_{t \to \infty} \frac{m^{-1}\ell(e_j^T A) + td_j m^{-1}\ell'(e_j^T A) - \bar{L}(A)}{t} \\ &= d_j m^{-1}\ell'(e_j^T A);\end{aligned}$$

which is positive by the selection of d_j and since $\ell' > 0$.

(Item 2 $\Rightarrow$ item 3.) Choose any $\ell \in L_{fs}$. Since the infimum is attainable, designate any satisfying $\inf \bar{L}(A) = \bar{L}(A)$ (note, although $\bar{L}$ is strictly convex, $\bar{L}$ A need not be, thus uniqueness is not guaranteed!). By the optimality conditions on the duality rule in Theorem 2.2.1, $_A = \ell'(A)/m > 0$ everywhere.

(Item 3 $\Rightarrow$ item 4.) This holds since $D_A \{ _A\}$ and $_A \in \mathbb{R}^m_{++}$ for every $\ell \in L_{fs}$.

(Item 4 =) item 1.) This case is directly handled by Stiemke's theorem (Borwein and Lewis, 2000, Exercise 2.2.8). □

2.4.3 General Setting

So far, the scenarios of weak learnability and attainability corresponded to the extremal hard core cases of $|C| \in \{0, m\}$. The situation in the general setting $1 \leq |C| \leq m-1$ is basically as good as one could hope for: it interpolates between the two extremal cases.

As a first step, partition A into two submatrices according to C.

Definition 2.4.6. Let $A \in \mathbb{R}^{m \times n}$ and corresponding hard core C be given. Partition A by rows into two matrices $A_0 \in \mathbb{R}^{m_0 \times n}$ and $A_+ \in \mathbb{R}^{m_+ \times n}$, where A_+ has rows corresponding to C, and $m_+ = |C|$. For convenience, permute the examples so that

$$A = \begin{bmatrix} A_0 \\ A_+ \end{bmatrix}.$$

(This merely relabels the coordinate axes, and does not change the optimization problem.) Note that this decomposition is unique, since C is uniquely specified.

As a first consequence, this partition cleanly decomposes the dual feasible set D_A into D_{A_0} and D_{A_+}.

Proposition 2.4.7. For any $A \in \mathbb{R}^{m \times n}$, $D_{A_0} = \{0\}$, $D_{A_+} \cap \mathbb{R}^{m_+}_{++} \neq \emptyset$, and

$$D_A = D_{A_0} \times D_{A_+}.$$

Furthermore, no other partition of A into $B_0 \in \mathbb{R}^{z \times n}$ and $B_+ \in \mathbb{R}^{p \times n}$ satisfies these properties.

Proof. It must hold that $D_{A_0} = \{0\}$, since otherwise there would exist $\lambda \in \ker(A_0^{\geq}) \cap \mathbb{R}^{m_0}_+$ with $\lambda \neq 0$, which could be extended to $\lambda' = \begin{bmatrix} \lambda \\ 0 \end{bmatrix} \in D_A$ and the positive coordinate of λ could be added to C, contradicting the maximality of C.

The property $D_{A_+} \cap \mathbb{R}^{m_+}_{++} \neq \emptyset$ is a consequence of Theorem 2.4.4 and the definition of C.

For the decomposition, note first that certainly every $\lambda \in D_{A_0} \times D_{A_+}$ satisfies $\lambda \in D_A$. Now suppose contradictorily that there exists $\lambda' \in D_A \setminus (D_{A_0} \times D_{A_+})$. There must exist $j \in [m] \setminus C$ with $(\lambda')_j > 0$, since otherwise $\lambda' \in \{0\} \times D_{A_+}$; but that means j should have been included in C, a contradiction.

Lastly, the uniqueness is a consequence of the uniqueness of hard cores over $\mathbb{R}^m$ (cf. Theorem 1.7.4, where uniqueness follows from the structure of $\mathcal{H}$). □

The main result of this section will have the same two main ingredients as Proposition 2.4.7:

> The full boosting instance may be uniquely decomposed into two pieces A_0 and A_+, each of which individually behave like the weak learnability and attainability scenarios.
>
> The subinstances have a somewhat independent effect on the full instance.

Theorem 2.4.8. Let $A \in \mathbb{R}^{m \times n}$ be given. Let $B_0 \in \mathbb{R}^{z \times n}$, $B_+ \in \mathbb{R}^{p \times n}$ be any partition of A by rows. The following conditions are equivalent.

1. $\exists\ \bar{x} \in \mathbb{R}^n$ $B_0 \bar{x} \in \mathbb{R}^z_- \wedge B_+ \bar{x} = 0$ and $\forall\ \bar{x} \in \mathbb{R}^n$ $B_+ \bar{x} \in \mathbb{R}^p_+ \setminus \{0\}$,
2. $\forall \grave{\ell} \in L_{fs}$, $\hat{L}(A, \grave{\ell}) = \hat{L}(B_+, \grave{\ell})$, and $\hat{L}(B_0, \grave{\ell}) = 0$, and $\hat{L}(\cdot, \grave{\ell}) + \iota_{\mathrm{im}(B_+)}$ is 0-coercive,
3. $\forall \grave{\ell} \in L_{fs}$, $\grave{\ell}_A = \begin{pmatrix} \grave{\ell}_{B_0} \\ \grave{\ell}_{B_+} \end{pmatrix}$ with $\grave{\ell}_{B_0} = 0$ and $\grave{\ell}_{B_+} \in \mathbb{R}^p_{++}$,
4. $D_{B_0} = 0$, and $D_{B_+} \setminus \mathbb{R}^p_{++} \neq \emptyset$, and $D_A = D_{B_0} \cap D_{B_+}$.

Stepping through these properties, notice that item 4 mirrors the expression in Proposition 2.4.7. But that proposition also granted that this representation was unique, thus only one partition of A satisfies the above properties, namely A_0, A_+. Since this theorem is stated as a series of equivalences, any one of these properties can in turn be used to identify the hard core set C.

To continue with geometric interpretations, notice that item 1 states that there exists a halfspace strictly containing those points in $[m] \cap C$, with all points of C on its boundary; furthermore, trying to adjust this halfspace to contain elements of C will place others outside it. With regards to the geometry of the dual feasible set as provided by item 4, the origin is within the relative interior of the points corresponding to C, however the convex hull of the other $m - |C|$ points can not contain the origin. Furthermore, if the origin is written as a convex combination of all points, this combination must place zero weight on the points with indices $[m] \cap C$. This scenario is depicted in Figure 2.3.

In items 2 and 3, B_0 mirrors the behavior of weakly learnable instances in Theorem 2.4.1, and analogously B_+ follows instances with minimizers from Theorem 2.4.4. The interesting addition, as discussed above, is the independence of these components: item 2 provides that the

Figure 2.3. Geometric view of the primal and dual problem in the general case. There is a closed homogeneous halfspace containing the points $\{a_i\}_{i=1}^m$, where the hard core lies on the halfspace boundary, and the other points are within its interior; moreover, there does not exist a closed homogeneous halfspace containing all points but with strict containment on a point in the hard core. Finally, although the origin is in the convex hull of $\{a_i\}_{i=1}^m$, any such convex combination places zero weight on points outside the hard core. Please see Theorem 2.4.8 and its discussion.

in mum of the combined problem is the sum of the in ma of the subproblems, while item 3 provides that the full dual optimum may be obtained by concatenating the subproblems' dual optima.

Proof of Theorem 2.4.8. (Item 1 =) item 2.) Let $\grave{} \in L_{fs}$ and be given with $B_0 \in R^z$ and $B_+ = 0$, and let $\{c_i\}_{1=1}^1 " 1$ be an arbitrary sequence increasing without bound. Lastly, let $\{ _i\}_1^1$ be a minimizing sequence for $\inf \L\b(B_+)$. Then

$$\begin{aligned} \inf m_+ \L\b(B_+) &= \lim_{i!1} \quad m_+ \L\b(B_+ _i) + m_0 \L\b(c_i B_0) \qquad \inf m \L\b(A) \\ &= \inf (m_+ \L\b(B_+) + m_0 \L\b(B_0)) \quad \inf m_+ \L\b(B_+); \end{aligned}$$

which used the fact that $\L\b(B_0) \quad 0$ since $\L\b \quad 0$. Since the chain of inequalities starts and ends the same, it must be a chain of equalities, which means $\inf \L\b(B_0) = 0$. To show 0-coercivity of $\L\b + _{im(B_+)}$, note the second part of item 1 is one of the conditions of Theorem 2.4.4.

(Item 2 =) item 3.) First, by Theorem 2.4.1, $\inf \L\b(B_0) = 0$ means $\grave{}_{B_0} = 0$ and

$D_{B_0} = \{0\}$. Thus

$$\begin{aligned} m_L(\hat{\ell}_A) &= \sup_{z \in D_A} m_L(z) \\ &= \sup\left\{ m_0(L_0(z_0)) + m_+(L_+(z_+)) : z_0 \in R^{z_0}_+, z_+ \in R^{p}_+, B_0^\top z_0 + B_+^\top z_+ = 0 \right\} \\ &\le \sup_{z_0 \in D_{B_0}} m_0(L_0(z_0)) + \sup_{z_+ \in D_{B_+}} m_+(L_+(z_+)) \\ &= \inf_{z \in R^n} m_+(L_+(B_+ z)) = \inf_{z \in R^n} m_L(Az) = m_L(\hat{\ell}_A). \end{aligned}$$

Combining this with $L(z) = \int_0^z \ell(b)\,db$ and $\ell(0) = 0$ (cf. Proposition 1.B.2 Lemma 1.B.1), $m_L(\hat{\ell}_A) = m_+(L_+(\hat{\ell}_{B_+})) = m_L(L(\hat{z}_{B_0}, \hat{z}_{B_+}))$. But Theorem 2.2.1 shows $\hat{\ell}_A$ was unique, which gives the result. To obtain $\hat{\ell}_{B_+} \in R^p_{++}$, use Theorem 2.4.4 with the 0-coercivity of $L_+ + \iota_{\mathrm{im}(B_+)}$.

(Item 3 $\Rightarrow$ item 4.) Since $\hat{\ell}_{B_0} = 0$, it follows by Theorem 2.4.1 that $D_{B_0} = \{0\}$. Furthermore, since $\hat{\ell}_{B_+} \in R^p_{++}$, it follows that $D_{B_+} \cap R^p_{++} \neq \emptyset$. Now suppose contradictorily that $D_A \neq D_{B_0} \times D_{B_+}$; since it always holds that $D_A \subseteq D_{B_0} \times D_{B_+}$, this supposition grants the existence of $z = \begin{bmatrix} z_0 \\ z_+ \end{bmatrix} \in D_A$ where $z_0 \in R^{z_0}_+ \setminus 0$.

Consider the element $q := z + \hat{\ell}_A$, which has more nonzero entries than $\hat{\ell}_A$, but still $q \in D_A$ since D_A is a convex cone. Let I_q index the nonzero entries of q, and let A_q be the restriction of A to the rows I_q. Since $q \in D_A$, meaning q is nonnegative and $q \in \ker(A^\top)$, it follows that the restriction of q to its positive entries is within $\ker(A_q^\top)$ (because only zeros of q and matching rows of A are removed, dot products between q with rows of $A^\top$ are the same as dot products between the restriction of q and rows of $A_q^\top$), and so $q \in D_{A_q}$, meaning $D_{A_q} \cap R^{|I_q|}_{++}$ is nonempty. Correspondingly, by Theorem 2.4.4, the dual optimum $\hat{\ell}_{A_q}$ of this restricted problem will have only positive entries. But by the same reasoning granting that q restricted to I_q is within D_{A_q}, it follows that the full optimum $\hat{\ell}_A$, restricted to I_q, must also be within D_{A_q} (since, by q's construction, $\hat{\ell}_A$'s zero entries are a superset of the zero entries of q). Therefore this restriction $\hat{\ell}_A$ of $\hat{\ell}_A$ to I_q will have at least one zero entry, meaning it can not be equal to $\hat{\ell}_{A_q}$; but Theorem 2.2.1 provided that the dual optimum is unique, thus $L(\hat{\ell}_{A_q}) > L(\hat{\ell}_A)$. Finally, produce $\hat{\ell}_{A_q}$ from $\hat{\ell}_{A_q}$ by inserting a zero for each entry of I_q; the same reasoning that allows feasibility to be maintained while removing zeros allows them to be added, and thus $\hat{\ell}_{A_q} \in D_A$. But this is a contradiction: since $\ell(0) = 0$ (cf. Lemma 1.B.1), both $\hat{\ell}_{A_q}$ and the optimum $\hat{\ell}_A$

have zero contribution to the objective along the entries outside of $\mathcal{I}_q$, and thus

$$m\underline{\ell}(\grave{A}_q) = \sum_{i \in \mathcal{I}_q} \underline{\ell}(\grave{A}_q e_i) > \sum_{i \in \mathcal{I}_q} \underline{\ell}(\hat{\grave{A}} e_i) = m\underline{\ell}(\hat{\grave{A}}),$$

meaning $\grave{A}_q$ is feasible and has strictly greater objective value than the optimum $\hat{\grave{A}}$.

(Item 4 $\Rightarrow$ item 1.) Unwrapping the definition of D_A, the assumed statements imply

$$(\forall \lambda_0 \in \mathbb{R}^z_+ \setminus \{0\}, \lambda_+ \in \mathbb{R}^p_+ : B_0^\top \lambda_0 + B_+^\top \lambda_+ \geq 0) \wedge (\exists \lambda_+ \in \mathbb{R}^p_{++} : B_+^\top \lambda_+ = 0).$$

Applying Motzkin's transposition theorem (cf. (Ben-Israel, 2002) or (Dantzig and Thapa, 2003, Theorem 2.16)) to the left statement and Stiemke's theorem (cf. (Borwein and Lewis, 2000, Exercise 2.2.8), which is implied by Motzkin's theorem) to the right yields

$$(\exists \lambda \in \mathbb{R}^n : B_0 \lambda \in \mathbb{R}^z \wedge B_+ \lambda \in \mathbb{R}^p) \wedge (\forall \lambda \in \mathbb{R}^n : B_+ \lambda \in \mathbb{R}^p \setminus \{0\}),$$

which implies the desired statement. □

Remark 2.4.9. Notice the dominant role A and corresponding hard core C play in the structure of the optimization solutions. For every $i \in [m] \setminus C$, the corresponding dual weights go to zero (i.e., $(r \underline{\ell}(A_t))_i \to 0$), and the corresponding primal margins grow unboundedly (i.e., $e_i^\top A_t \to \infty$, since otherwise $\inf \underline{\ell}(A_0 \cdot) > 0$). This is completely unaffected by the choice of $\grave{} \in L_{fs}$. Furthermore, whether this instance is weak learnable, attainable, or neither is dictated purely by A (respectively $|C| = 0$, $|C| = m$, or $|C| \in [1; m-1]$).

Where different loss functions disagree is how they assign dual weight to the points in C. In particular, each $\grave{} \in L_{fs}$ (and corresponding $\underline{\ell}$) defines a notion of entropy via $\underline{\ell}$. The dual optimization in Theorem 2.2.1 can then be interpreted as selecting the max entropy choice (per $\underline{\ell}$) amongst those convex combinations of C equal to the origin.

Appendix 2.A Generalizing the Weak Learning Rate

2.A.1 Choosing a Generalization to $\gamma(b)$

Any generalization γ' of $\gamma(b)$ should satisfy the following properties.

When empirical weak learnability holds (i.e., when $\gamma(b) > 0$), then $\gamma' = \gamma(b)$.

For any boosting instance, $\gamma^0 \in (0, 1]$.

γ^0 provides an expression similar to eq. (2.3.5), which allows the full gradient to be converted into a notion of suboptimality in the dual.

Taking the form of the classical weak learning rate from eq. (2.3.1) as a model, the template generalized weak learning rate is

$$\gamma^0(A, S, C, D) := \inf_{\lambda \in S \cap C} \frac{\|A^\top \lambda\|_1}{\inf_{\lambda \in S \setminus D} \|\lambda\|_1},$$

for some sets S, C, and D (for instance, the classical weak learning rate uses $S = \mathbb{R}^m_+$ and $C = D = \{0\}$). In order to provide an expression similar to eq. (2.3.5), the domain of the infimum must include every suboptimal dual iterate $\lambda \in L^0(A, \lambda_t)$.

Any choice C which does not include all of $\ker(A^\top)$ is immediately problematic: this allows $\lambda \in S \setminus \ker(A^\top)$ to be selected, whereby $\|A^\top \lambda\| = 0$ and $\gamma^0 = 0$. But note that without being careful about D, it is still possible to force the value 0.

Remark 2.A.1. Another generalization is to define

$$\gamma^{00}(A) := \gamma^0(A, \mathbb{R}^m_+, \ker(A^\top), \{\bar{\lambda}_A\}) = \inf_{\lambda \in \mathbb{R}^m_+ \cap D_A} \frac{\|A^\top \lambda\|_1}{\|\lambda - \bar{\lambda}_A\|_1}. \tag{2.A.2}$$

This form agrees with the original when weak learnability holds, and leads to a very convenient analog to eq. (2.3.5).

Unfortunately, γ^{00} may be zero. Specifically, take A to be the matrix S defined in Section 3.4, due to Schapire (2010), where

$$\bar{\lambda}_S = \frac{\ell'(0)}{3} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Furthermore, for any $\alpha \in (0, 1)$, define

$$\lambda_\alpha := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \in \operatorname{im}(S), \qquad \lambda := (1 - \alpha) \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix} + \alpha \bar{\lambda}_S \in \ker(S^\top).$$

Then

$$\inf_{\lambda \in \mathbb{R}^m_+ \cap \ker(S^\top)} \frac{\|S^\top \lambda\|_1}{\|\lambda - \bar{\lambda}_S\|_1} \leq \inf_{\alpha \in (0,1)} \frac{\|S^\top (\alpha \lambda_\alpha + (1 - \alpha) \lambda)\|_1}{\|\alpha \lambda_\alpha + (1 - \alpha) \lambda - \bar{\lambda}_S\|_1} = \inf_{\alpha \in (0,1)} \frac{1 - \alpha}{1} = 0.$$

The natural correction to these worries is to set $C = D = \ker(A^>)$. But there is still sensitivity due to S.

Remark 2.A.3. Set $A := (1;1)^>$, meaning $\ker(A^>) = \{z(1;-1)^> : z \in \mathbb{R}\}$, and $S = B((1;1)^>; \sqrt{2})$, the ball of radius $\sqrt{2}$ around $(1;1)^>$; note that $S \cap \ker(A^>) = 0$. Consider $\theta'(A; S; \ker(A^>); \ker(A^>))$, and the sequence $\{x_i\}_{i=1}^{1}$ where

$$x_i = (1;1)^> - \frac{1}{\sqrt{i^2+1}} \begin{bmatrix} 2 \\ i+1 \end{bmatrix} \begin{bmatrix} 3 \\ i \end{bmatrix} \begin{bmatrix} 7 \\ -1 \end{bmatrix}.$$

Note that $\|x_i - (1;1)^>\|_2 = \sqrt{2}$, thus $x_i \in S$. Furthermore, $A^> x_i \neq 0$, so $x_i \notin S \cap \ker(A^>)$. As such,

$$\theta'(A; S; \ker(A^>); \ker(A^>)) \leq \inf_i \frac{\|Ax_i\|_1}{\|x_i - P^1_{S\cap\ker(A^>)}(x_i)\|_1} \tag{2.A.4}$$

$$= \frac{\|(1;1)^> - (1;1)\sqrt{i^2+1}\begin{bmatrix} i+1 \\ -1 \end{bmatrix}\|_1}{\|(1;1)_2 - \sqrt{i^2+1}\begin{bmatrix} i+1 \\ -1 \end{bmatrix}\|_1}. \tag{2.A.5}$$

Using $\sqrt{y} \leq (1+y)/2$, the numerator has upper bound

$$\|(1;1)^> - (1;1)\sqrt{i^2+1}\begin{bmatrix} i+1 \\ -1 \end{bmatrix}\|_1 = |2\sqrt{i^2+1} - 2i|$$
$$= 2i(\sqrt{1+i^{-2}} - 1)$$
$$\leq 2i((2+i^{-2})/2 - 1) = 1/i.$$

The denominator is

$$\|(1;1)^> - \sqrt{i^2+1}\begin{bmatrix} i+1 \\ -1 \end{bmatrix}\|_1 = |\sqrt{i^2+1} - (i+1)| + |\sqrt{i^2+1} - (i-1)|$$
$$= ((i+1) - \sqrt{i^2+1}) + (\sqrt{i^2+1} - (i-1))$$
$$= 2.$$

Thus eq. (2.A.5) is bounded above by $\inf_i (2i)^{-1} = 0$.

The difficulty here was the curvature of S, which allowed elements arbitrarily close to $\ker(A^>)$ without actually being inside this subspace. This possibility is averted here by requiring

polyhedrality of S. This choice is sufficiently rich to allow the various dual-distance upper bounds of Chapter 3.

2.A.2 Proof of Theorem 2.3.6

The proof of Theorem 2.3.6 requires a few steps, but the strategy is straightforward. First note that $(A;S)$ can be rewritten as

$$\begin{aligned}
(A;S) &= \inf_{\in S\cap\ker(A^\top)} \frac{\|A^\top \|_1}{\| - P^1_{S\cap\ker(A^\top)}( )\|_1} \\
&= \inf_{\in S\cap\ker(A^\top)} \frac{\|A^\top (- P^1_{S\cap\ker(A^\top)}( ))\|_1}{\| - P^1_{S\cap\ker(A^\top)}( )\|_1} \\
&= \inf\left\{ \frac{\|A^\top v\|_1}{\|v\|_1} : v \in \mathbb{R}^m \setminus \{0\}, \exists \in S,\ v = - P^1_{S\cap\ker(A^\top)}( ) \right\} \\
&= \inf\left\{ \|A^\top v\|_1 : \|v\|_1 = 1, \exists \in S, \exists c > 0,\ cv = - P^1_{S\cap\ker(A^\top)}( ) \right\}; \qquad (2.A.6)
\end{aligned}$$

where the second equivalence used that $A^\top P^1_{S\cap\ker(A^\top)}( ) = 0$.

In the final form, $v \neq 0$ and $v \in \ker(A^\top)$, and so $A^\top v \neq 0$; that is to say, the infimand is positive for every element of its domain. The difficulty is that the domain of the infimum, written in this way, is not obviously closed; thus one cannot simply assert the infimum is attainable and positive.

The goal then will be to reparameterize the infimum to have a compact domain. For technical convenience, the result will be mainly proved for the l_2 norm (where projections behave nicely), and norm equivalence will provide the final result.

Lemma 2.A.7. Given $A \in \mathbb{R}^{m\times n}$ and a polyhedron $S \subseteq \mathbb{R}^m$ with $S \cap \ker(A^\top) \neq \emptyset$ and $S \cap \ker(A^\top) \neq \{0\}$,

$$\inf\left(\frac{\|A^\top (- P^2_{S\cap\ker(A^\top)}( ))\|_2}{\| - P^2_{S\cap\ker(A^\top)}( )\|_2} : \in S \cap \ker(A^\top) \right) > 0. \qquad (2.A.8)$$

To produce the desired reparameterization of this infimum, the following characterization of polyhedral sets will be used.

Definition 2.A.9. For any nonempty polyhedral set $S \subseteq \mathbb{R}^m$, let H_S index a finite (but possibly empty) collection of affine functions $g : \mathbb{R}^m \to \mathbb{R}$ so that $S = \cap_{\in H_S}\{x \in \mathbb{R}^m : g (x) \geq 0\}$ (with the convention that $S = \mathbb{R}^m$ when $H_S = \emptyset$). For any $x \in S$, let $I_S(x)$ denote the active set for x: $\in I_S(x)$ iff $g (x) = 0$. Lastly, define a relation $\preceq_S$ over points in S: given $x, y \in S$, $x \preceq_S y$ iff

$I_S(x) = I_S(y)$. Observe that $\sim_S$ is an equivalence relation over points within S, and let C_S be the set of equivalence classes.

The equivalence relation $\sim_S$ thus partitions S into the members of C_S, each of which has a very convenient structure.

Lemma 2.A.10. Let a polyhedral set $S \subseteq R^m$ be given, and x a nonempty $F \in C_S$. Then F is convex, and F is equal to its relative interior (i.e., $F = ri(F)$). Finally, fixing an arbitrary $z_0 \in F$, the normal cone at any point $z \in F$ is orthogonal to the vector space parallel to the affine hull of F (i.e., $N_F(z) = (aff(F) - z)^\perp = (aff(F) - z_0)^\perp$).

Throughout the remainder of this section, normal and tangent cones will be considered at points within a set $F \in C_S$. As Lemma 2.A.10 establishes, any set $F \in C_S$ is relatively open ($F = ri(F)$), however, the required properties of normal and tangent cones, as developed by Hiriart-Urruty and Lemaréchal (2001, Sections A.5.2 and A.5.3), suppose closed convex sets. But it is always the case that $ri(F) = ri(cl(F))$ (Hiriart-Urruty and Lemaréchal, 2001, Proposition A.2.1.8); as such, the normal and tangent cones at the desired relative interior points may just as well be constructed against $cl(F)$, and thus the aforementioned properties safely hold.

Proof. If $S = R^m$ (meaning H_S is empty) or $\dim(F) = 0$ (F is a single point), everything follows directly, thus suppose $S \neq R^m$, and x a nonempty $F \in C_S$ with $\dim(F) > 0$.

Let any $x_0, x_1 \in F$ and $\in [0,1]$ be given, and define $x := (1-)x_0 + x_1$. Since each g defining S is affine,

$$g(x) = (1-)g(x_0) + g(x_1). \tag{2.A.11}$$

By construction of C_S, $g(x_0) = 0$ iff $g(x_1) = 0$ and otherwise both are negative, thus $g(x) = 0$ iff $g(x_0) = g(x_1) = 0$, meaning $I_S(x) = I_S(x_0) = I_S(x_1)$, so $x \in F$ and F is convex.

Now let any $y_0 \in F$ be given; $y_0 \in ri(F)$ when there exists > 0 so that

$$B(y_0,) \cap aff(F) \subseteq F \tag{2.A.12}$$

(Hiriart-Urruty and Lemaréchal, 2001, Definition A.2.1.1). To this end, first define to be half the distance to the closest hyperplane defining S which is not active for y_0:

$$:= \frac{1}{2} \min_{\in H_S \cap I_S(y)} \min\{\|y' - y_0\|_2 : y' \in R^m, g(y') = 0\}.$$

Since there are only finitely many such hyperplanes, and the distance to each is nonzero, $\epsilon > 0$. Let any $y \in B(y_0, \epsilon) \cap \text{aff}(F)$ be given; by definition of $\text{aff}(F)$, there must exist $\alpha \in \mathbb{R}$ and $y_1 \in F$ so that $y = (1-\alpha)y_0 + \alpha y_1$. By eq. (2.A.11), for any $i \in I_S(y_0) = I_S(y_1)$,

$$g_i(y) = (1-\alpha)g_i(y_0) + \alpha g_i(y_1) = 0.$$

On the other hand, for any $i \in H_S \cap I_S(y_0)$, it must be the case that $g_i(y) < 0$, since $y \in B(y_0, \epsilon)$, and due to the choice of ϵ. Returning to the definition of relative interior in eq. (2.A.12), it follows that $y_0 \in \text{ri}(F)$, and $\text{ri}(F) = F$ since $y_0 \in F$ was arbitrary.

For the final property, for any $z_0, z \in \text{ri}(F) = F$, the tangent cone $T_F(z)$ has form $(\text{aff}(F) - z)$ (see Hiriart-Urruty and Lemaréchal, 2001, Proposition A.5.2.1 and discussion within Section A.5.3), and note $\text{aff}(F) - z = \text{aff}(F) + z_0 - z - z_0 = \text{aff}(F) - z_0$. Lastly, $N_F(z) = T_F(z)^*$ (Hiriart-Urruty and Lemaréchal, 2001, Proposition A.5.2.4). □

The relevance to eq. (2.A.8) and eq. (2.A.6) is that projections from polyhedron S onto $S \cap \ker(A^\top)$ (itself a polyhedron, as is verified in the proof of Lemma 2.A.7) must land on some equivalence class of $C_{S \cap \ker(A^\top)}$, and these projections are easily characterized.

Lemma 2.A.13. Let any nonempty polyhedra $S \subseteq \mathbb{R}^m$ and $K \subseteq \mathbb{R}^m$ be given, and x any nonempty $F \in C_{S \cap K}$ and $x_F \in F$. Define

$$P_F := \{c(\Pi^2_{S \cap K}(\gamma)) : c > 0,\ \gamma \in S, \Pi^2_{S \cap K}(\gamma) \in F\},$$
$$D_F := N_F(x_F) \cap \{-\gamma x_F : \gamma \in \mathbb{R}^m, \forall i \in I_S(x_F),\ g_i(\gamma) \geq 0\},$$

where $N_F(x_F)$ is the normal cone of F at x_F. Then $P_F = D_F$.

Note that the final active set $I_S(x_F)$ is with respect to S, not $S \cap K$.

Proof. ($\subseteq$) Let any $\gamma \in S$ with $\alpha := \Pi^2_{S \cap K}(\gamma) \in F$ be given, where the latter is well-defined since F and hence $S \cap K$ are nonempty. By Lemma 2.A.10, $\alpha \in \text{ri}(F)$, and $N_F(\alpha) = N_F(x_F)$, meaning $-\gamma x_F \in N_F(x_F)$ (Hiriart-Urruty and Lemaréchal, 2001, Proposition A.5.3.3). Since $\gamma \in S$, for

any $2 I_S(\ ) = I_S(x_F)\ \ H_S, g(\ )\ \ 0$, so

$$2 f y 2 R^m : g(y)\ \ 0g\ f\ \ g = f y 2 R^m : g(y)\ \ 0g\ f\ \ x_F g\ \ f\ x_F g$$
$$= f y 2 R^m : g(y)\ \ 0g\ f\ x_F g;$$

the nal equality following since $g(x_F) = g(\ ) = 0$ and g de nes an a ne hyperplane, meaning the corresponding a ne halfspace is closed under translations by x_F. This holds for all $2 I_S(x_F)$, thus $2 D_F$, and since D_F is a convex cone, for any $c > 0$, $c(\ ) 2 D_F$.

( ) De ne

$$:= \min fk x_F\ \ zk_2 :\ \ 2 H_S n I_S(x_F); z 2 R^m; g(z) = 0 g:$$

For any xed , this minimum is positive since $g(x_F) < 0$, while polyhedrality of S grants that ranges over a nite set, together meaning > 0. Now let any $v 2 D_F$ be given, and set $:= x_F + v=(2kvk_2)$. The form of D_F immediately grants $g(\ )\ \ 0$ for $2 I_S(x_F)$, but notice for $2 H_S n I_S(x_F)$, it still holds that $g(\ )\ \ 0$, since $g(x_F) < 0$ and $k\ \ x_F k_2 < $. So $v = (2 kvk_2=)(\ \ P^2_{S\setminus K}(\ ))$ where $2 S$ and $P^2_{S\setminus K}(\ ) = x_F\ 2 F$, meaning $v 2 P_F$. □

The result now follows by considering all elements of $C_{S\setminus \ker(A^>)}$.

Proof of Lemma 2.A.7. For convenience, set $K := \ker(A^>)$. Note that K (and hence $S \setminus K$) is a polyhedron; indeed, it has the form

$$K = \ker(A^>) = f\ \ 2 R^m : A^>\ \ = 0 g$$
$$= \bigcap_{i=1}^{n} f\ \ 2 R^m : e_i^> A^>\ \ \ 0g \setminus f\ \ 2 R^m : e_i^> A^>\ \ \ 0g\ :$$

Next, note $C_{S\setminus K}$ has at least one nonempty equivalence class, since $S \setminus K$ is nonempty by assumption. Rewriting eq. (2.A.8) as in eq. (2.A.6), and xing an x_F within each nonempty

$F \in C_{S\setminus K}$, Lemma 2.A.13 grants

$$\begin{aligned}(2.A.8) &= \inf \{\|A^\top v\|_2 : \|v\|_2 = 1, \exists c > 0, \exists \in S \quad P^2_{S\setminus K}(\ ) = cv\} \\ &= \min_{\substack{F \in C_{S\setminus K} \\ F \neq \{0\}}} \inf \{\|A^\top v\|_2 : \|v\|_2 = 1, \exists c > 0, \exists \in S \quad P^2_{S\setminus K}(\ ) = cv, P^2_{S\setminus K}(\ ) \in F\} \\ &= \min_{\substack{F \in C_{S\setminus K} \\ F \neq \{0\}}} \inf \{\|A^\top v\|_2 : \|v\|_2 = 1, v \in N_F(x_F), \forall \in I_S(x_F) \ \ g(x_F + v) \ \ 0\}.\end{aligned}$$

Since $S \cap \ker(A^\top) \neq \{0\}$, and $S \setminus \ker(A^\top)$, at least one infimum has a nonempty domain (for the others, take the convention that their value is $+1$). Each infimum with a nonempty domain in this final expression is of a continuous function over a compact set (in fact, a polyhedral cone intersected with the boundary of the unit l^2 ball), and thus it has a minimizer v, which corresponds to some $c(\ P^2_{S\setminus K}(\ )) \in \ker(A^\top)$, where $c > 0$. It follows that

$$A^\top v = cA^\top(\ P^2_{S\setminus K}(\ )) \neq 0,$$

meaning each of these infima is positive. But since S is polyhedral, C_S has finitely many equivalence classes ($|C_S| \ \ 2^{|H_S|}$), meaning the outer minimum is attained and positive. □

Finally, as mentioned above, the desired result follows by norm equivalence.

Proof of Theorem 2.3.6. For the upper bound, note as in the proof of Lemma 2.A.7 that $S \setminus \ker(A^\top) \neq \{0\}$; and the infimand is positive for every element of the domain, so the infimum is finite. For the lower bound, by Lemma 2.A.7 and norm equivalence,

$$\begin{aligned}(A, S) = &\inf_{\in S \cap \ker(A^\top)} \frac{\|A^\top\|_{k_1}}{\inf_{\in S\setminus\ker(A^\top)} \|\|_{k_1}} \\ &\frac{1}{\sqrt{mn}} \inf_{\in S \cap \ker(A^\top)} \frac{\|A^\top\|_{k_2}}{\inf_{\in S\setminus\ker(A^\top)} \|\|_{k_2}} > 0.\end{aligned}$$

□

Appendix 2.B Proof of 0-coercivity Characterization

Proof of Proposition 2.4.3. Suppose a minimizer x exists, and define

$$S_b := \{x + v : \|v\|_2 = b\},$$
$$r_b := \inf\{h(x) : x \in S_b\}.$$

For any $b > 0$, it cannot hold that $r_b < h(x)$, since then there exists x with $h(x) < h(x)$; furthermore, it also can not hold that $r_b = h(x)$, since the closure property on h guarantees there exists x_b with $x_b \in S_b$ and $h(x_b) = r_b$, and by strict convexity, the point $x' := (x_b + x)/2$ has $h(x') < h(x)$. Next, convexity grants $2(r_b - h(x)) \geq r_{2b} - h(x)$, since for any x_2 with $\|x_2 - x\|_2 = 2b$,

$$r_b - h(x) \leq h((x_2 + x)/2) - h(x) < (h(x_2) - h(x))/2;$$

since x_2 was arbitrary, it holds for the infimum over all such x_2, which gives the inequality. It follows that all sublevel sets are compact.

On the other hand, if all sublevel sets are compact, since h is closed and proper, a minimizer exists. □

Appendix 2.C Bibliographic Notes

Bibliographic notes on hard cores appear in Section 1.E. Hints of the dual problem may be found in many works, most notably those of Kivinen and Warmuth (1999), Collins et al. (2002), which demonstrated that boosting is seeking a difficult distribution over training examples via iterated Bregman projections.

Acknowledgements

This chapter is based on work by the dissertation author which appeared in the Journal of Machine Learning Research, Volume 13, pages 561–606, 2012.

Chapter 3

Optimization Guarantees

This chapter develops rates of numerical convergence of AdaBoost under the setting of Chapter 2; namely, only a finite sample is considered (meaning the measure is $\hat{\mu}$ and the surrogate risk is $\hat{L}$), and the image set $\{(h(x_i))_{i=1}^m : h \in H\}$ is assumed finite, whereby A may simply be written as a matrix $A \in \mathbb{R}^{m \times n}$.

Looking forward, if H is simply finite, then the numerical convergence rates here can be combined with the statistical guarantees of Chapter 5. Developing a statistical analysis from the rates here for infinite H seems challenging, however, as the bounds here grow quickly with the sample size m. Consequently, Chapter 6, which provides guarantees for infinite H, will provide a separate numerical analysis; note however that the numerical rates there are much slower than the ones here.

Chapter 6 and the present chapter not only disagree on the size of H which they can handle; the later chapter elects to handle only certain Lipschitz losses, for instance the logistic loss $\ell_{\log}$. Interestingly, that analysis fails in a number of places for the exponential loss, and meanwhile the analysis here has terrible constants growing quickly with the sample size in the case of the logistic loss. As such, the mechanisms in this thesis are still not adequately sensitive to these losses.

3.1 Overview

Convergence rates will be proved for the following family of loss functions, which refines L_{fs} by requiring some derivative bounds over any level set.

Definition 3.1.1. L_{fso} contains all functions ℓ satisfying the following properties. First, $\ell \in L_{fs}$. Second, for any $x \in \mathbb{R}^m$ satisfying $\hat{L}(x) - \hat{L}(A_0) = \ell(0)$, and for any coordinate $(x)_i$, there exist

constants > 0 and > 0 such that $`'(x)_i) \quad `(x)_i)$ and $`(x)_i) \quad `'(x)_i)$.

The exponential loss $`_{exp}$ is in this family with $= = 1$ since $\exp(\ )$ is a xed point with respect to the di erentiation operator. Furthermore, as is veri ed in Remark 3.A.1, the logistic loss $`_{log}$ is also in this family, with $= 2^m=(m \ln(2))$ and $= 1 + 2^m$ (which may be loose). In a sense, and encode how similar some $` 2 L_{fso}$ is to the exponential loss, and thus these parameters can degrade radically. However, outside the weak learnability case, the other terms in the bounds here can also incur a large penalty with the exponential loss, and there is some evidence that this is unavoidable (see the lower bounds in Mukherjee et al. (2011) or the upper bounds in Rätsch et al. (2001)).

The rates are presented in three separate sections, mirroring the optimization structure in Chapter 2. First, Sections 3.2 and 3.3 respectively consider weak learnable and attainable instances (i.e., instances where the hard core is respectively empty or contains every example), which entails that the rate is $O(\ln(1=))$. The general case is presented in Section 3.4, where the rate degrades to $O(1=)$ (with a matching lower bound for $`_{log}$); this degradation is due to the preceding cases being fast for very di erent reasons, and as will be discussed after the proof, these di erent reasons come into con ict in the general case. In each section, there will be some discussion on how to slightly weaken the conditions on L_{fso}, however in each case this will lead to degraded rates. For example, the relationship between function values and derivative values in the de nition of L_{fso} may seem unusual, however it will be the key to the $O(\ln(1=))$ rate under the empirical weak learning assumption.

Each proof will use the following basic guarantee on the progress granted by a single step of the algorithm. This fact mostly follows from standard guarantees of line search methods, however it also leverage basic properties of $(A; S)$, exactly as presented in eq. (2.3.5); namely, $(A; S)$ allows $A^>$ to be removed from a normed expression.

Proposition 3.1.2. For any t, $` 2 L_{fso}$, $A 2 R^{m \ n}$, and $S \ fr \ L(A \ _t)g$ with $(A; S) > 0$,

$$L(A \ _{t+1}) \quad L(A \ ) \quad L(A \ _t) \quad L(A \ ) \quad \frac{^2 \ (A; S)^2 D^1_{S \setminus \ker(A^>)}(r \ L(A \ _t))^2}{6 \ L(A \ _t)}:$$

Proof of Proposition 3.1.2. If $_t$ is optimal, then $r \ L(A \ _t) = 0$, and moreover since every line search option performs descent, $_{t+1} = \ _t$, whereby the desired guarantee holds with equality (potentially using the convention $0=1$ in the second nal term). As such, suppose in the remainder

of the proof that $_t$ is not optimal.

Proceeding as in eq. (2.3.4), since S fr $\hat{L}(A_t)g$,

$$(A;S) = \inf_{2 S\cap\ker(A^>)} \frac{kA^> \quad k_1}{D^1_{S\setminus\ker(A^>)}(\ )} \quad \frac{kA^> r\hat{L}(A_t)k_1}{D^1_{S\setminus\ker(A^>)}(r\hat{L}(A_t))}:$$

By Lemma 3.A.2, $\hat{L}$ A has Lipschitz gradients with parameter $\hat{L}(A_t)$ with respect to $L^1(b)$ over the $(t+1)^{st}$ level set, whereby the preconditions for the guarantees for all three line searches in Lemmas 1.C.3 and 1.C.7 are met, yielding

$$\hat{L}(A_t) \quad \hat{L}(A_{t+1}) \quad \frac{^2kA^> r\hat{L}(A_t)k_1^2}{6\ \hat{L}(A_t)} \quad \frac{^2\ (A;S)^2D^1_{S\setminus\ker(A^>)}(r\hat{L}(A_t))^2}{6\ \hat{L}(A_t)}:$$

Subtracting $\hat{L}(A\ )$ from both sides and rearranging yields the statement. □

From here, the three di erent cases will manage the dual distance $D^1_{S\setminus\ker(A^>)}(r\hat{L}(A_t))$, speci cally producing a relation to $\hat{L}(A_t) \quad \hat{L}(A\ )$, the total suboptimality in the preceding iteration; from there, standard tools in convex optimization will yield convergence rates. Matching the problem structure revealed in Section 2.4, rst the extremal cases of weak learnability and attainability will be handled, and only then the general case.

3.2 Weak Learnability

First consider the case that the empirical weak learning assumption holds, which implies the dual set D_A is exactly $f0g$ by Proposition 2.3.2, which in turn implies the hard core C is empty.

Theorem 3.2.1. Suppose $A\ 2\ R^{m\ \ n}$, and the corresponding hard core C satis es $jCj = 0$. Let any $` 2 L_{fso}$ be given. Then $(A;R^m_+) > 0$, and for any $t \quad 0$,

$$\hat{L}(A_t) \quad \hat{L}(A_0) \quad 1 \quad \frac{^2\ (A;R^m_+)^2}{6\ ^2} \quad ^t:$$

Proof. By Theorem 2.4.1, $D_A = f0g$, meaning

$$D^1_{D_A}(r\hat{L}(A_t)) = \inf_{2D_A} kr\hat{L}(A_t) \quad k_1 = kr\hat{L}(A_t)k_1 \quad \hat{L}(A_t)=\ :$$

Next, R_+^m is polyhedral, and Theorem 2.4.1 grants $R_+^m \cap \ker(A^\top) \neq \{0\}$; and $R_+^m \cap \ker(A^\top) \neq \{0\}$, so Theorem 2.3.6 provides $\gamma(A; R_+^m) > 0$. Since $\hat{L}(A_t) \in R_+^m$, all conditions of Proposition 3.1.2 are met, and using $\hat{L}(A_1) = 0$ (again by Theorem 2.4.1),

$$\hat{L}(A_{t+1}) - \hat{L}(A_t) \leq \frac{\eta^2 \gamma(A; R_+^m)^2 \hat{L}(A_t)^2}{8\eta^2 \hat{L}(A_t)} = \hat{L}(A_t)\left(1 - \frac{\eta^2 \gamma(A; R_+^m)^2}{8\eta^2}\right); \tag{3.2.2}$$

and recursively applying this inequality yields the result. □

Specializing this analysis to the exponential loss (where $\eta = \eta = 1$), assuming that a best choice $h_t \in \mathcal{H}$ exists (whereby $\eta = 1$), and recalling from Section 2.3 that $\gamma(A; R_+^m) = \gamma(b)$ (the empirical weak learning rate), the bound becomes $(1 - \gamma(b)^2/8)^t$. This bound in turn recovers the bound due to Schapire and Singer (1999), albeit with the denominator degraded from 2 to 8 (which can be recovered by choosing only step size options 1 & 2); this is despite the analysis here going through a very different path.

In general, still making use of $\gamma(A; R_+^m) = \gamma(b)$, solving for t in the expression

$$\epsilon = \frac{\hat{L}(A_t) - \hat{L}(A)}{\hat{L}(A_0) - \hat{L}(A)} \leq \left(1 - \frac{\eta^2 \gamma(b)^2}{8\eta^2}\right)^t \leq \exp\left(-\frac{t\eta^2 \gamma(b)^2}{8\eta^2}\right)$$

reveals that $t \geq \frac{8\eta^2}{\eta^2 \gamma(b)^2} \ln(1/\epsilon)$ iterations suffice to reach suboptimality ϵ. Recall that η and η, in the case of the logistic loss, have only been bounded by quantities like m^2. While it is unclear if this analysis of η and η was tight, note that it is plausible that the logistic loss is slower than the exponential loss in this scenario, as it works less in initial phases to correct minor margin violations. As stated previously, however, it may simply be the case that the mechanism used here to simultaneously analyze the logistic and exponential losses is simply insensitive to the logistic loss.

Remark 3.2.3. The rate $O(\ln(1/\epsilon))$ depended crucially on both $\ell' < \ell'(0)$ and $\ell'' < \ell'$. If for instance the second inequality were replaced with $\ell'' \leq C$, then eq. (3.2.2) would instead have form $\hat{L}(A_{t+1}) - \hat{L}(A_t) \leq \hat{L}(A_t)^2 O(1)$, which by an application of standard convex optimization tools would grant a rate $O(1/\epsilon)$ (Shalev-Shwartz and Singer, 2008, Lemma 20). For functions which asymptote to zero (i.e., everything in $\mathcal{L}_{fs}$), satisfying this milder second order condition is quite easy. The real mechanism behind producing a fast rate is $\ell' < \ell'(0)$, which guarantees that the flattening of the objective function is concomitant with low objective values.

3.3 Attainability

Consider now the case of attainability. Recall from Theorem 2.4.4 and Proposition 2.4.3 that attainability occurred along with a stronger property: the function $\bar{b} + \iota_{\mathrm{im}(A)}$ is 0-coercive, meaning it has compact level sets. Recall further that it was necessary to restrict $\bar{b}$ to the slice im(A) via $\bar{b} + \iota_{\mathrm{im}(A)}$, since A may have linearly dependent columns, which immediately destroys any hope of compact level sets.

This level set structure has an immediate consequence to the task of relating $\bar{b}(A_t) - \bar{b}(A)$ to the dual distance $D^1_{S \setminus \ker(A^T)}(r\bar{b}(A_t))$. $\bar{b}$ is a strictly convex function, which means it is strongly convex over any compact set. Strong convexity in the primal corresponds to upper bounds on second derivatives in the dual (occasionally termed *strong smoothness*), which in turn can be used to relate distance and objective values. Making these statements rigorous will also provide the choice of polyhedron S in (A, S): unlike the case of weak learnability which chose $S = \mathbb{R}^m_+$, a compact set containing the initial level set will be used.

Theorem 3.3.1. *Suppose $A \in \mathbb{R}^{m \times n}$, and the corresponding hard core C satisfies $|C| = m$. Let any $\ell \in L_{fso}$ be given. Then there exists a (compact) tightest axis-aligned rectangle C containing the initial level set $\{x \in \mathbb{R}^m : (\bar{b} + \iota_{\mathrm{im}(A)})(x) \le \bar{b}(A_0)\}$, and $\bar{b}$ is strongly convex with modulus $c > 0$ over C. Finally, $(A, r\bar{b}(C)) > 0$, and for all $\epsilon > 0$ and t,*

$$\bar{b}(A_t) - \bar{b}(A) \le (\bar{b}(A_0) - \bar{b}(A)) \left(1 - \frac{c^2 (A, r\bar{b}(C))^2}{3\bar{b}(A_0)}\right)^t.$$

As in Section 3.2, when ϵ_0 is suboptimal, this bound may be rearranged to say that $t \ge \frac{3\bar{b}(A_0)}{c^2 (A, r\bar{b}(C))^2} \ln(1/\epsilon)$ iterations suffice to reach suboptimality ϵ.

To make sense of this bound and its proof, the essential object is C, whose properties are captured in the following lemma, which is stated with some slight generality in order to allow reuse in Section 3.4.

Lemma 3.3.2. *Let $A \in \mathbb{R}^{m \times n}$, corresponding hard core C with $|C| = m$, any $\ell \in L_{fso}$, and any $\delta \ge \inf \bar{b}(A)$ be given. Then there exists a (compact nonempty) tightest axis-aligned rectangle $C \supseteq \{x \in \mathbb{R}^m : (\bar{b} + \iota_{\mathrm{im}(A)})(x) \le \delta\}$. Furthermore, the dual image $r\bar{b}(C) \subseteq \mathbb{R}^m$ is also a (compact nonempty) axis-aligned rectangle, and moreover $r\bar{b}(C) \subseteq \mathrm{int}(\mathrm{dom}(\bar{b}^*)) \subseteq \mathbb{R}^m_{++}$. Finally, $r\bar{b}(C)$ contains dual feasible points (i.e., $r\bar{b}(C) \cap D_A \neq \emptyset$).*

A full proof may be found in Section 3.A.3; the principle is that $|C| = m$ provides 0-coercivity of $L^* + \delta_{\mathrm{im}(A)}$, and thus the initial level set is compact. To later show $\sigma(A, S) > 0$ via Theorem 2.3.6, S must be polyhedral, and to apply Proposition 3.1.2, it must contain the dual iterates $\{\nabla L^*(A x_t)\}_{t=1}^{\infty}$; the easiest choice then is to take the bounding box C of the initial level set, and use its dual map $\nabla L^*(C)$. To exhibit dual feasible points within $\nabla L^*(C)$, note that C will contain a primal minimizer, and optimality conditions grant that $\nabla L^*(C)$ contains the dual optimum.

With the polyhedron in place, Proposition 3.1.2 may be applied, so what remains is to control the dual distance. Again, this result will be stated with some extra generality in order to allow reuse in Section 3.4.

Lemma 3.3.3. Let $A \in \mathbb{R}^{m \times n}$, any $\ell \in \mathcal{L}_{\mathrm{fso}}$, and any compact set S with $\nabla L^*(S) \cap \ker(A^*) \neq \emptyset$ be given. Then L^* is strongly convex over S, and taking $c > 0$ to be the modulus of strong convexity, for any $x \in S \cap \mathrm{im}(A)$,

$$L^*(x) - L^*(A\bar{x}) \leq \frac{1}{2c} \inf_{z \in \nabla L^*(S) \cap \ker(A^*)} \|z - \nabla L^*(x)\|_1^2.$$

Before presenting the proof, it can be sketched quite easily. Using the Fenchel-Young inequality and the dual optimization problem (cf. Theorem 2.2.1), primal suboptimality can be converted into a Bregman divergence in the dual. If there is strong convexity in the primal, it allows this Bregman divergence to be converted into a distance via standard tools in convex optimization (Shalev-Shwartz and Singer, 2008, Lemma 18). Although L^* lacks strong convexity in general, it is strongly convex over any compact set.

Proof of Lemma 3.3.3. Consider the optimization problem

$$\inf_{x \in S} \inf_{\substack{k \in \mathbb{R}^m \\ \|k\|_2 = 1}} \left\langle \nabla^2 L^*(x), k \right\rangle = m^{-1} \inf_{x \in S} \inf_{\substack{k \in \mathbb{R}^m \\ \|k\|_2 = 1}} \sum_{i=1}^{m} \ell''(x_i) \langle e_i, k \rangle^2,$$

since S is compact and ℓ'' and $(\cdot)^2$ are continuous, the infimum is attained. But $\ell'' > 0$ and $S \neq \emptyset$, meaning the infimum c is nonzero, and moreover it is the modulus of strong convexity of L^* over S (Hiriart-Urruty and Lemaréchal, 2001, Theorem B.4.3.1.iii).

Now let any $x \in S \cap \mathrm{im}(A)$ be given, define $D = \nabla L^*(S) - \mathbb{R}^m_+$, and for convenience set $K := \ker(A^*)$. Consider the dual element $P^2_{D \cap K}(\nabla L^*(x))$ (which exists since $D \cap K \neq \emptyset$); due to

the projection, it is dual feasible, and thus it must follow from Theorem 2.2.1 and the form of ψ in Proposition 1.B.2 that

$$\psi(A\,) = \sup\{\, \psi\,(\ ) :\ \ 2\ D_A\} \qquad \psi\ \ P^2_{D\setminus K}(r\ \psi(x))\ \ :$$

Furthermore, since $x \in \mathrm{im}(A)$,

$$\left\langle x;\, P^2_{D\setminus K}(r\ \psi(x))\right\rangle = 0:$$

Combined with the Fenchel-Young inequality and the general property $x = r\ \psi\,(r\ \psi(x))$ (Rockafellar, 1970, Theorem 23.5),

$$\begin{aligned}
\psi(x)\ \ \psi(A\,) &\le \psi(x) + \psi\ \ P^2_{D\setminus K}(r\ \psi(x)) \\
&= \psi\ \ P^2_{D\setminus K}(r\ \psi(x)) \ \ + \ \ \left\langle r\ \psi(x); x\right\rangle \qquad \psi\,(r\ \psi(x)) \\
&= \psi\ \ P^2_{D\setminus K}(r\ \psi(x)) \qquad \psi\,(r\ \psi(x)) \qquad \left\langle r\ \psi\,(r\ \psi(x)); P^2_{D\setminus K}(r\ \psi(x))\ r\ \ \psi(x)\right\rangle \qquad (3.3.4)\\
&\le \frac{1}{2c}\|r\ \psi(x) \ \ P^2_{D\setminus K}(r\ \psi(x))\|_2^2; \qquad (3.3.5)
\end{aligned}$$

where the last step follows by a property of Bregman divergences in the dual (Shalev-Shwartz and Singer, 2008, Lemma 18), noting that both $r\ \psi(x)$ and $P^2_{D\setminus K}(r\ \psi(x))$ are in $r\ \psi(S) = D$, and ψ is strongly convex with modulus c over S. To nish, rewrite P as an in mum and use $\|\ \|_2 \ \ \|\ \|_1$. □

The desired result now follows readily.

Proof of Theorem 3.3.1. Invoking Lemma 3.3.2 with $d = \psi(A_{\ 0})$ immediately provides a compact tightest axis-aligned rectangle C containing the initial level set $S := \{x \in R^m : (\psi + _{\mathrm{im}(A)})(x) \le \psi(A_{\ 0})\}$. Crucially, since the objective values never increase, S and C contain every iterate $\{A_{\ t}\}_{t=1}^{1}$.

Applying Lemma 3.3.3 to the set C (by Lemma 3.3.2, $r\ \psi(C) \cap \ker(A^{>}) \neq \ ;$), then for any t,

$$\psi(A_{\ t}) \ \ \psi(A\,) \qquad \frac{1}{2c}\|r\ \psi(A_{\ t}) \ \ P^1_{r\ \psi(C)\setminus \ker(A^{>})}(r\ \psi(A_{\ t}))\|_1^2;$$

where $c > 0$ is the modulus of strong convexity of ψ over C.

Finally, if there are suboptimal iterates, then $r\,\mathcal{L}_b(C) \neq r\,\mathcal{L}_b(S)$ contains points that are not dual feasible, meaning $r\,\mathcal{L}_b(C) \cap \ker(A^\top) \neq \emptyset$; since Lemma 3.3.2 also provided $r\,\mathcal{L}_b(C) \cap D_A \neq \emptyset$, and $r\,\mathcal{L}_b(C)$ is a hypercube, it follows by Theorem 2.3.6 that $\delta(A, r\,\mathcal{L}_b(C)) > 0$. Plugging this into Proposition 3.1.2 and using $\mathcal{L}_b(A_{t}) \leq \mathcal{L}_b(A_0)$ gives

$$\begin{aligned}\mathcal{L}_b(A_{t+1}) - \mathcal{L}_b(A) &\leq \mathcal{L}_b(A_t) - \mathcal{L}_b(A) - \frac{\delta^2(A, r\,\mathcal{L}_b(C))^2 D^{-1}_{r\,\mathcal{L}_b(C)\cap\ker(A^\top)}\,(r\,\mathcal{L}_b(A_t))^2}{6\,\mathcal{L}_b(A_t)} \\ &\leq (\mathcal{L}_b(A_t) - \mathcal{L}_b(A))\left(1 - \frac{c^2\,\delta^2(A, r\,\mathcal{L}_b(C))^2}{3\,\mathcal{L}_b(A_0)}\right);\end{aligned}$$

and the result again follows by recursively applying this inequality. □

Remark 3.3.6. The key conditions on $\ell \in \mathcal{L}_{\mathrm{fso}}$, namely the existence of constants granting $\ell \leq \ell'$ and $\ell'' \leq \ell$ within the initial level set, are much more than are needed in this setting. Inspecting the presented proofs, it entirely suffices that on any compact set in $\mathbb{R}^m$, $\mathcal{L}_b$ has quadratic upper and lower bounds (equivalently, bounds on the smallest and largest eigenvalues of the Hessian), which are precisely the weaker conditions used in previous treatments (Bickel et al., 2006, Rätsch et al., 2001).

These quantities are therefore necessary for controlling convergence under weak learnability. To see how the proofs of this section break down in that setting, consider the central Bregman divergence expression in eq. (3.3.4). What is really granted by attainability is that every iterate lies well within the interior of $\mathrm{dom}(\mathcal{L}_b)$, and therefore these Bregman divergences, which depend on $r\,\mathcal{L}_b$, can not become too wild. On the other hand, with weak learnability, all dual weights go to zero (cf. Theorem 2.4.1), which means that $r\,\ell \to 1$, and thus the upper bound in eq. (3.3.5) ceases to be valid. As such, another mechanism is required to control this scenario, which is precisely the role of $\ell \leq \ell'$ and $\ell'' \leq \ell$.

3.4 General Setting

The key development of Section 2.4.3 was that general instances may be decomposed uniquely into two smaller pieces, one satisfying attainability and the other satisfying weak learnability, and that these smaller problems behave somewhat independently. This independence is leveraged here to produce convergence rates relying upon the existing rate analysis for the attainable and weak learnable cases. The mechanism of the proof is as straightforward as one

could hope for: decompose the dual distance into the two pieces, handle them separately using preceding results, and then stitch them back together.

Theorem 3.4.1. Let $A \in \mathbb{R}^{m \times n}$, corresponding hard core C with $1 \le |C| \le m-1$, and any loss $\ell \in \mathcal{L}_{fso}$ be given. Recall from Section 2.4.3 the partition of the rows of A into $A_0 \in \mathbb{R}^{m_0 \times n}$ and $A_+ \in \mathbb{R}^{m_+ \times n}$, and suppose the axes of $\mathbb{R}^m$ are ordered so that $A = \begin{bmatrix} A_0 \\ A_+ \end{bmatrix}$. Set C_+ to be the tightest axis-aligned rectangle containing $C_+ := \{x \in \mathbb{R}^{m_+} : (L^\ell + \iota_{\mathrm{im}(A_+)})(x) \le L^\ell(A_0)\}$, and $w := \sup_t \|t - L^\ell(A_+ t) - P^1_{r L^\ell(C_+) \cap \ker(A_+^\ge)}(r - L^\ell(A_+ t))\|_1$. Then C_+ is compact, $w < 1$, L^ℓ has modulus of strong convexity $c > 0$ over C_+, and $\gamma(A, R^{m_0} - r L^\ell(C_+)) > 0$. Using these terms, for all t,

$$L^\ell(A t) - L^\ell(A) \le \frac{2L^\ell(A_0)}{(t+1)\min\left\{1, \gamma^2(A, R_+^{m_0} - r L^\ell(C_+))^2/(3(1 + w/(2c))^2)\right\}}.$$

The new term, w, appears when stitching together the two subproblems. For choices of $\ell \in \mathcal{L}_{fso}$ where $\mathrm{dom}(\ell^*)$ is a compact set, this value is easy to bound; for instance, the logistic loss, where $\mathrm{dom}(L^{\ell^*}) = [0, 1/m]^m$, has $w \le \sup_{\alpha \in \mathrm{dom}(L^{\ell^*})} \|\alpha\|_1 = 1$ (since $0 \in \mathrm{dom}(L^{\ell^*})$). And with the exponential loss, taking $S := \{\alpha \in \mathbb{R}^n : L^\ell(A\alpha) \le L^\ell(A_0)\}$ to denote the initial level set, since 0 is always dual feasible,

$$w \le \sup_{\alpha \in S} \|r - L^{\ell^*}(A\alpha)\|_1 = \sup_{\alpha \in S} 1 - L^{\ell^*}(A\alpha) = 1 - L^{\ell^*}(A_0) = 1.$$

Note that rearranging the rate from Theorem 3.4.1 will provide that $O(1/\epsilon)$ iterations suffice to reach suboptimality ϵ, whereas the earlier scenarios needed only $O(\ln(1/\epsilon))$ iterations. The exact location of the degradation will be pinpointed after the proof, and is related to the introduction of w.

Proof of Theorem 3.4.1. By Theorem 2.4.8, $m_+ L^\ell(A_+ \alpha) = m L^\ell(A\alpha)$, and the form of L^ℓ gives $m L^\ell(A t) = m_0 L^\ell(A_0 t) + m_+ L^\ell(A_+ t)$, thus

$$m L^\ell(A t) - m L^\ell(A\alpha) = m_0 L^\ell(A_0 t) + m_+ L^\ell(A_+ t) - m_+ L^\ell(A_+ \alpha). \tag{3.4.2}$$

For the left term, since $\ell(x) \ge |\ell'(x)|$,

$$L^\ell(A_0 t) - \|r - L^\ell(A_0 t)\|_1 = \|r - L^\ell(A_0 t) - P^1_{A_0}(r - L^\ell(A_0 t))\|_1; \tag{3.4.3}$$

which used the fact (from Theorem 2.4.8) that $\mathcal{A}_0 = \{0\}$.

For the right term of eq. (3.4.2), recall from Theorem 2.4.8 that $\hat{b} + \mathrm{im}(A_+)$ is 0-coercive, thus the level set $S_+ := \{x \in \mathbb{R}^{m_+} : (\hat{b} + \mathrm{im}(A_+))(x) \leq \hat{b}(A_0)\}$ is compact. For all t, since $\hat{b} \geq 0$ and the objective values never increase,

$$m\hat{b}(A_0) \leq m\hat{b}(A_t) = m_0\hat{b}(A_{0\,t}) + m_+\hat{b}(A_{+\,t}) \leq m_+\hat{b}(A_{+\,t});$$

in particular, $A_{+\,t} \in S_+$. It is crucial that the level set compares against $\hat{b}(A_0)$ and not $\hat{b}(A_{+\,0})$.

Continuing, Lemma 3.3.2 may be applied to A_+ with value $d = \hat{b}(A_0)$, which grants a tightest axis-aligned rectangle $C_+ \supseteq \mathbb{R}^{m_+}$ containing S_+, and moreover $r\,\hat{b}(C_+) \cap \ker(A_+^>) \neq \emptyset$. Applying Lemma 3.3.3 to A_+ and C_+, $\hat{b}$ is strongly convex with modulus $c > 0$ over C_+, and for any t,

$$\hat{b}(A_{+\,t}) - \hat{b}(A_+) \leq \frac{1}{2c}\|r\,\hat{b}(A_{+\,t}) - P^1_{r\,\hat{b}(C_+)\cap\ker(A_+^>)}(r\,\hat{b}(A_{+\,t}))\|_1^2. \tag{3.4.4}$$

Next, set $w := \sup_t \|r\,\hat{b}(A_{+\,t}) - P^1_{r\,\hat{b}(C_+)\cap\ker(A^>)}(r\,\hat{b}(A_{+\,t}))\|_1$; $w < 1$ since S_+ is compact and $r\,\hat{b}(C_+) \cap \ker(A^>)$ is nonempty. By the definition of w,

$$D^1_{r\,\hat{b}(C_+)\cap\ker(A_+^>)}(r\,\hat{b}(A_{+\,t}))^2 \leq wD^1_{r\,\hat{b}(C_+)\cap\ker(A_+^>)}(r\,\hat{b}(A_{+\,t})),$$

which combined with eq. (3.4.4) yields

$$\hat{b}(A_{+\,t}) - \hat{b}(A_+) \leq \frac{w}{2c}D^1_{r\,\hat{b}(C_+)\cap\ker(A_+^>)}(r\,\hat{b}(A_{+\,t})). \tag{3.4.5}$$

To merge the subproblem dual distance upper bounds eq. (3.4.3) and eq. (3.4.5) via Lemma 3.A.3, it must be shown that $(\mathbb{R}^{m_0}_+ \times r\,\hat{b}(C_+)) \cap D_A \neq \emptyset$. But this follows by construction and Theorem 2.4.8, since $\{0\} = \mathcal{A}_0 \times \mathbb{R}^m_+ \times r\,\hat{b}(C_+) \cap \mathcal{A}_+ \neq \emptyset$ by Lemma 3.3.2, and the decomposition $D_A = \mathcal{A}_0 \times \mathcal{A}_+$. Returning to the total suboptimality expression eq. (3.4.2),

these dual distance bounds yield

$$
\begin{aligned}
m\hat{\mathcal{L}}(A\lambda_t) - m\hat{\mathcal{L}}(A\bar\lambda) &\leq m_0\, D^1_{A_0}(\nabla\hat{\mathcal{L}}(A_0\lambda_t)) + m_+ w/(2c) D^1_{\nabla\hat{\mathcal{L}}(C_+)\cap\ker(A_+^\top)}(\nabla\hat{\mathcal{L}}(A_+\lambda_t)) \\
&\leq m(\eta + w/(2c))\, D^1_{(\mathbb{R}^{m_0}_+\times\partial\hat{\mathcal{L}}(C_+))\cap\ker(A^\top)}(\nabla\hat{\mathcal{L}}(A\lambda_t));
\end{aligned}
$$

the second step using Lemma 3.A.3.

To finish, note $\mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+)$ is polyhedral, and

$$
(\mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+)) \setminus \ker(A^\top) \supseteq \{\nabla\hat{\mathcal{L}}(A\lambda_t)\}_{t=1}^{\infty} \setminus \ker(A^\top) \neq \emptyset;
$$

since no primal iterate is optimal and thus $\nabla\hat{\mathcal{L}}(A\lambda_t)$ is not dual feasible by optimality conditions; combined with the above derivation $(\mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+)) \cap D_A \neq \emptyset$, Theorem 2.3.6 may be applied, meaning $\gamma(A; \mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+)) > 0$. As such, all conditions of Proposition 3.1.2 are met, and making use of $\hat{\mathcal{L}}(A\lambda_t) \leq \hat{\mathcal{L}}(A\lambda_0)$,

$$
\begin{aligned}
\hat{\mathcal{L}}(A\lambda_{t+1}) - \hat{\mathcal{L}}(A\bar\lambda) &\leq \hat{\mathcal{L}}(A\lambda_t) - \hat{\mathcal{L}}(A\bar\lambda) \\
&\quad - \frac{\rho^2\gamma(A; \mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+))^2 D^1_{(\mathbb{R}^{m_0}_+\times\partial\hat{\mathcal{L}}(C_+))\cap\ker(A^\top)}(\nabla\hat{\mathcal{L}}(A\lambda_t))^2}{6\eta\hat{\mathcal{L}}(A\lambda_t)} \\
&\leq \hat{\mathcal{L}}(A\lambda_t) - \hat{\mathcal{L}}(A\bar\lambda) - \frac{\rho^2\gamma(A; \mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+))^2(\hat{\mathcal{L}}(A\lambda_t) - \hat{\mathcal{L}}(A\bar\lambda))^2}{6\eta\hat{\mathcal{L}}(A\lambda_0)(\eta + w/(2c))^2}.
\end{aligned}
$$

Applying a standard conversion from convex optimization, for instance as stated by Shalev-Shwartz and Singer (2008, Lemma 20) and invoking it with parameters

$$
\epsilon_t := \frac{\hat{\mathcal{L}}(A\lambda_t) - \hat{\mathcal{L}}(A\bar\lambda)}{\hat{\mathcal{L}}(A\lambda_0)} \qquad \text{and} \qquad r := \frac{1}{2}\min\left\{1; \frac{\rho^2\gamma(A; \mathbb{R}^{m_0}_+ \times \partial\hat{\mathcal{L}}(C_+))^2}{3\eta(\eta + w/(2c))^2}\right\};
$$

the result follows. □

In order to produce a rate $O(\ln(1/\epsilon))$ under attainability, strong convexity related the suboptimality to a *squared* dual distance $\|\cdot\|_1^2$ (cf. eq. (3.3.5)). On the other hand, the rate $O(\ln(1/\epsilon))$ under weak learnability came from a fortuitous cancellation with the denominator $\hat{\mathcal{L}}(A\lambda_t)$ (cf. eq. (3.2.2)), which is equal to the total suboptimality since Theorem 2.4.1 provides $\hat{\mathcal{L}}(A\bar\lambda) = 0$. But in order to merge the subproblem dual distances via Lemma 3.A.3, the differing

properties granting fast rates must be ignored. (In the case of attainability, this process introduces w.)

This incompatibility is not merely an artifact of the analysis. Intuitively, the nite and in nite margins sought by the two pieces $A_0; A_+$ are in con ict. For a beautifully simple, concrete case of this, consider the following matrix, due to Schapire (2010):

$$S := \begin{bmatrix} 2 & 3 \\ 1 & +1 \\ +1 & 1 \\ 1 & 1 \end{bmatrix}:$$

The optimal solution here is to push both coordinates of unboundedly positive, with margins approaching $(0; 0; 1)$. But pushing any coordinate $( )_i$ too quickly will increase the objective value, rather than decreasing it. In fact, this instance will provide a lower bound, and the mechanism of the proof shows that the primal weights grow extremely slowly, as $\Omega(\ln(t))$.

Theorem 3.4.6. Consider $`_{log} = \ln(1 + \exp( ))$ 2 L_{fso}, the logistic loss, and suppose the line search is exact (cf. option 1 in Figure 1.3). Then for any t 1, $L(S_t)$ $L(S)$ $1=(24t)$.

Finally, note that this third setting does not always entail slow convergence. Again taking the view of the rows of S being points $f s_i g_{i=1}^3$, consider the e ect of rotating the entire instance around the origin by $= 4$. The optimization scenario is unchanged, however coordinate descent can now be arbitrarily close to the optimum in one iteration by pushing a single primal weight extremely high.

Appendix 3.A Miscellaneous Technical Material

3.A.1 The Logistic Loss is within L_{fso}

Remark 3.A.1. This remark develops bounds on the quantities ; for the logistic loss $`_{log} = \ln(1 + \exp( ))$. First note that the initial level set $S_0 := f x 2 R^m : L(x)$ $L(A_0)g$ is contained within a cube $( 1 ; b]^m$, where $b$ $m \ln(2)$; this follows since $L(A_0) = L(0) = \ln(2)$, whereas $`_{log}(m \ln(2)) = \ln(1 + \exp(m \ln(2))) > m \ln(2)$.

For convenience, the analysis will be mainly written with respect to $b = m \ln(2)$. Let any $x 2 (1 ; b]$ be given, and note $`'_{log} = \exp( )=(1 + \exp( ))$, and $`''_{log} = \exp( )=(1 + \exp( ))^2$.

To determine , note 1 $1 + \exp(x)$ $1 + \exp(b)$. Since $\ln$ is concave, it follows for all

$z \in [1; 1 + \exp(b)]$ that the secant line through $(1; 0)$ and $(1 + \exp(b); \ln(1 + \exp(b)))$ is a lower bound:

$$\ln(z) \quad \frac{\ln(1 + \exp(b)) \quad 0}{1 + \exp(b) \quad 1} \quad z \quad \frac{\ln(1 + \exp(b)) \quad 0}{1 + \exp(b) \quad 1} = \ln(1 + \exp(\ b)) \exp(\quad b)(z \quad 1):$$

As such, for $x \in (\ 1 \quad ; b]$, $\ln(1 + \exp(x)) \quad \exp(x) \ln(1 + \exp(\ b)) \exp(\quad b)$, so

$$\frac{\grave{}''_{\log}(x)}{\grave{}_{\log}(x)} = \frac{\exp(x)}{(1 + \exp(x))^2 \ln(1 + \exp(x))} \quad \frac{\exp(b)}{(1 + \exp(x))^2 \ln(1 + \exp(\ b))} \quad \frac{\exp(b)}{\ln(1 + \exp(\ b))}:$$

Consequently, a su cient choice is $:= \exp(\ b) = \ln(1 + \exp(\ b)) \quad 2^m = (m \ln(2))$.

For $\grave{}_{\log}(x) \quad \grave{}'_{\log}(x)$, using $\ln(x) \quad x \quad 1$,

$$\frac{\grave{}_{\log}(x)}{\grave{}'_{\log}(x)} = \frac{\ln(1 + \exp(x))}{\frac{\exp(x)}{1+\exp(\ x)}} \quad \frac{\exp(x)}{\frac{\exp(x)}{1+\exp(\ x)}} \quad 1 + \exp(b):$$

That is, it su ces to set $:= 1 + \exp(\ b) = 1 + 2^{\ m}$.

3.A.2 L_{fso} implies L^b has Lipschitz Gradients

Lemma 3.A.2. Let $\grave{} \in L_{fso}$, any H (in nite is ne), and any empirical measure b over $X \quad f \quad 1; +1g$ be given. Then $L^b \quad A$ has Lipschitz gradients with parameter $L^b(A_{\ t \ 1})$ with respect to norm $L^1(b)$ over the t^{th} level set

$$S_t := \{ \quad 2 \ : \ L^b(A \ ) \quad L^b(A_{\ t \ 1}) \ \}:$$

Proof. Starting from Lemma 1.C.2, for any $; \ ^0 \in S_t$, since $\grave{}'' \quad 0$ (whereby $j\grave{}''j = \grave{}''$), and also

using the definition of the level set S_t,

$$\begin{aligned}
&\|\nabla(L^b_\mu A)(\lambda) - \nabla(L^b_\mu A)(\lambda')\|_1 \\
&\quad \frac{1}{m}\sum_{i=1}^m |\ell'((A\lambda)_{x_i,y_i}) - \ell'((A\lambda')_{x_i,y_i})| \\
&= \frac{1}{m}\sum_{i=1}^m \int_0^1 \ell''((A\lambda' + r(A\lambda - A\lambda'))_{x_i,y_i})(A\lambda - A\lambda')_{x_i,y_i}\, dr \\
&\quad \int_0^1 \frac{1}{m}\sum_{i=1}^m \ell''((A\lambda' + r(A\lambda - A\lambda'))_{x_i,y_i})dr \max_j y_j \sum_{h\in H} h(x_j)(\mu(h) - \mu'(h)) \\
&\quad L^b(A\lambda_{t-1})\|\lambda - \lambda'\|_1
\end{aligned}$$

as desired. □

3.A.3 Proof of Lemma 3.3.2

Proof of Lemma 3.3.2. Since $d > \inf_\lambda L^b(A\lambda)$, the level set $S_d := \{x \in \mathbb{R}^m : (L^b_\mu + \iota_{\mathrm{im}(A)})(x) \le d\}$ is nonempty. Since $|C| = m$, Theorem 2.4.4 provides $L^b_\mu + \iota_{\mathrm{im}(A)}$ is 0-coercive, meaning S_d is compact.

Now consider the rectangle C defined as a product of intervals $C = \prod_{i=1}^m [a_i, b_i]$, where

$$a_i := \inf\{x_i : x \in S_d\}, \qquad b_i := \sup\{x_i : x \in S_d\}.$$

By construction, $C \supseteq S_d$, and furthermore any smaller axis-aligned rectangle must violate some infimum or supremum above, and so must fail to include a piece of S_d. In particular, the tightest rectangle exists, and it is C.

Next, note that $\nabla L^b(x) = (\ell'(x_1)/m, \ell'(x_2)/m, \ldots, \ell'(x_m)/m)$, thus $D = \prod_{i=1}^m \ell'([a_i, b_i])/m$, an axis-aligned rectangle in the dual. Since ℓ is strictly convex and $\mathrm{dom}(\ell) = \mathbb{R}$, both $\ell'(a_i)$ and $\ell'(b_i)$ are within $\mathrm{int}(\mathrm{dom}(\ell^*))$ (for all i), and so $\nabla L^b(C) \subseteq \mathrm{int}(\mathrm{dom}(L^{b*}))$.

Finally, Proposition 2.4.3 grants that $L^b_\mu + \iota_{\mathrm{im}(A)}$ has a minimizer; thus choose any $\lambda \in \mathbb{R}^n$ so that $L^b(A\lambda) = \inf_\lambda L^b(A\lambda)$. By the optimality conditions in Theorem 2.2.1, $\lambda_A = \nabla L^b(A\lambda)$. But the dual optimum is dual feasible, and $\lambda_A \in S_d$, so

$$\nabla L^b(C) \cap D_A \supseteq \{\nabla L^b(A\lambda)\} \cap D_A = \{\lambda_A\} \cap D_A \neq \emptyset;$$

□

3.A.4 Splitting Distances along A_0, A_+

Lemma 3.A.3. Let $A = \begin{bmatrix} A_0 \\ A_+ \end{bmatrix}$ be given as in Theorem 3.4.1, and let a set $S = S_0 \cup S_+$ be given with $S_0 \subseteq R^{m_0}$ and $S_+ \subseteq R^{m_+}$ and $S \setminus D_A \neq \emptyset$. Then, for any $\lambda = \begin{bmatrix} \lambda_0 \\ \lambda_+ \end{bmatrix}$ with $\lambda_0 \in R^{m_0}$ and $\lambda_+ \in R^{m_+}$,

$$D^1_{S \setminus D_A}(\lambda) = D^1_{S_0 \setminus D_{A_0}}(\lambda_0) + D^1_{S_+ \setminus D_{A_+}}(\lambda_+).$$

Proof. Recall from Theorem 2.4.8 that $D_A = D_{A_0} \cup D_{A_+}$, thus

$$S \setminus D_A = (S_0 \setminus D_{A_0}) \cup (S_+ \setminus D_{A_+}),$$

and $S \setminus D_A \neq \emptyset$ grants that $S_0 \setminus D_{A_0} \neq \emptyset$ and $S_+ \setminus D_{A_+} \neq \emptyset$. Define now the notation $[\,]_0 : R^m \to R^{m_0}$ and $[\,]_+ : R^m \to R^{m_+}$, which respectively select the coordinates corresponding to the rows of A_0, and the rows of A_+.

Let $\lambda = \begin{bmatrix} \lambda_0 \\ \lambda_+ \end{bmatrix} \in R^m$ be given; in the above notation, $\lambda_0 = [\lambda]_0$ and $\lambda_+ = [\lambda]_+$. By the above Cartesian product and intersection properties,

$$\begin{bmatrix} P^1_{S_0 \setminus D_{A_0}}(\lambda_0) \\ P^1_{S_+ \setminus D_{A_+}}(\lambda_+) \end{bmatrix} \in S \setminus D_A,$$

and so

$$D^1_{S \setminus D_A}(\lambda) \begin{bmatrix} \lambda_0 \\ \lambda_+ \end{bmatrix} \begin{bmatrix} P^1_{S_0 \setminus D_{A_0}}(\lambda_0) \\ P^1_{S_+ \setminus D_{A_+}}(\lambda_+) \end{bmatrix}_1 = D^1_{S_0 \setminus D_{A_0}}(\lambda_0) + D^1_{S_+ \setminus D_{A_+}}(\lambda_+).$$

On the other hand, since $P^1_{S \setminus D_A}(\lambda) \in (S_0 \setminus A_0) \cup (S_+ \setminus A_+)$,

$$D^1_{S_0 \setminus D_{A_0}}(\lambda_0) + D^1_{S_+ \setminus D_{A_+}}(\lambda_+) \leq \|\lambda_0\|_1 + \|[P^1_{S \setminus D_A}(\lambda)]_0\|_1 + \|\lambda_+\|_1 + \|[P^1_{S \setminus D_A}(\lambda)]_+\|_1 = D^1_{S \setminus D_A}(\lambda).$$

□

3.A.5 Proof of Theorem 3.4.6

Proof of Theorem 3.4.6. This proof proceeds in two stages: first the gap between any solution with l_1 norm B is shown to be large, and then it is shown that the l_1 norm of the Boost solution (under logistic loss) grows slowly.

To start, $\ker(S^{>}) = \{z(1;1;0) : z \in R\}$, and $\ell_{\log}$ is maximized at $\ell^{0}_{\log}(0)$ with value $\ell_{\log}(0)$ (cf. Lemma 1.B.1). Thus $\ell_S = (\ell^{0}_{\log}(0); \ell^{0}_{\log}(0); 0) = 3$, and $L^b(S) = L^b(\ell_S) = 2\ell_{\log}(0) = 3 = 2\ln(2) = 3$.

Next, by calculus, given any B,

$$\inf_{k\,k_1\,B} L^b(S) \quad L^b(S) = L^b \quad S^{\left[\begin{smallmatrix} B=2 \\ B=2 \end{smallmatrix}\right]} \quad 2\ln(2) = 3$$

$$= (2\ln(2) + \ln(1 + \exp(\quad B))) = 3 \quad 2\ln(2) = 3$$

$$= \ln(1 + \exp(\quad B)) = 3:$$

Now to bound the l_1 norm of the iterates. By the nature of exact line search, the coordinates of are updated in alternation (with arbitrary initial choice); thus let u_t denote the value of the coordinate updated in iteration t, and v_t be the one which is held xed. (In particular, $v_t = u_{t\ 1}$.)

The objective function, written in terms of $(u_t; v_t)$, and ignoring the term $1=m = 1=3$ which will be deleted after di erentiation, is

$$\ln\ 1 + \exp(v_t \quad u_t) + \ln\ 1 + \exp(u_t \quad v_t) + \ln\ 1 + \exp(\quad u_t \quad v_t)$$

$$= \ln\ 2 + \exp(v_t \quad u_t) + \exp(u_t \quad v_t) + 2\exp(\quad u_t \quad v_t) + \exp(\quad 2u_t) + \exp(\quad 2v_t) :$$

Due to the use of exact line search, and the fact that u_t is the new value of the updated variable, the derivative with respect to u_t of the above expression must equal zero. In particular, producing this equality and multiplying both sides by the (nonzero) denominator yields

$$\exp(v_t \quad u_t) + \exp(u_t \quad v_t) \quad 2\exp(\quad u_t \quad v_t) \quad 2\exp(\quad 2u_t) = 0:$$

Multiplying by $\exp(u_t + v_t)$ and rearranging, it follows that, after line search, u_t and v_t must satisfy

$$\exp(2u_t) = \exp(2v_t) + 2\exp(v_t \quad u_t) + 2: \qquad (3.A.4)$$

First it will be shown for $t \quad 1$, by induction, that $u_t \quad v_t$. The base case follows by inspection (since $u_0 = v_0 = 0$ and so $u_1 = \ln(2)$). Now the inductive hypothesis grants $u_t \quad v_t$; the case $u_t = v_t$ can be directly handled by eq. (3.A.4), thus suppose $u_t > v_t$. But previously, it

was shown that the optimal l_1 bounded choice has both coordinates equal; as such, the current iterate, with coordinates $(u_t; v_t)$, is worse than the iterate $(u_t; u_t)$, and thus the line search will move in a positive direction, giving u_{t+1} v_{t+1}.

It will now be shown by induction that, for t 1, u_t $\frac{1}{2}\ln(4t)$. The base case follows by the direct inspection above. Applying the inductive hypothesis to the update rule above, and recalling $v_{t+1} = u_t$ and that the weights increase (i.e., u_{t+1} $v_{t+1} = u_t$),

$$\exp(2u_{t+1}) = \exp(2u_t) + 2\exp(u_t \quad u_{t+1}) + 2 \quad \exp(2u_t) + 2\exp(u_t \quad u_t) + 2 \quad 4t + 4 \quad 4(t+1):$$

To nish, recall by Taylor expansion that $\ln(1+q)$ q $\frac{q^2}{2}$; consequently for t 1

$$L(S_t) \quad L(S) \quad \inf_{k\ k_1\ \ln(4t)} L(S) \quad L(S) \quad \frac{1}{3}\ln\left(1 + \frac{1}{4t} \quad \frac{1}{12t} \quad \frac{1}{6}\left(\frac{1}{4t}\right)^2 \quad \frac{1}{24t}\right):$$

□

Appendix 3.B Bibliographic Notes

The study of general numerical convergence properties of AdaBoost variants was initiated by Collins et al. (2002), which provided convergence for a variety of step sizes and losses. The rst rate was $O(\exp(1/\varepsilon^2))$ to achieve accuracy $\varepsilon > 0$, proved in a very general setting by Bickel et al. (2006); this work later served as the basis for the rst consistency proof of AdaBoost, due to Bartlett and Traskin (2007).

In terms of rates under various assumptions, there is of course the original $O(\ln(1/\varepsilon)/\gamma(b)^2)$ rate due to Freund and Schapire (1997) when $\gamma(b) > 0$; a reasonable fast convergence rate was also shown with the logistic loss by Du y and Helmbold (2000). On the other end of the spectrum, when minimizers exist, Rätsch et al. (2001) established a fast (Q-linear) convergence rate using techniques due to Luo and Tseng (1993). As discussed in Chapter 2, hard cores identify these two cases as extremal (hard core is empty or everything), and the general case is a combination of the two.

Parallel to the original work behind this thesis chapter, Mukherjee et al. (2011) established general convergence under the exponential loss, with a rate of $\Theta(\frac{1}{\varepsilon})$. Just as with this chapter, their analysis split the problem into two pieces according to hard cores. Notably, they also proved

a $O(\ ^{5})$ iteration bound to beat a predictor of some xed norm. This second bound ends up being the essential one when controlling statistical properties in the general case, and a similar bound will be produced for the logistic loss (and similar losses) in Chapter 6, and moreover this bound was used to produce a streamlined consistency analysis of AdaBoost (Schapire and Freund, 2012, Theorem 12.2).

There are boosting variants with good convergence rates, for instance by Rätsch and Warmuth (2002), Warmuth et al. (2007), Shalev-Shwartz and Singer (2008). These variants, however, adjust the objective function, in particular adding regularization and constraints, and thus are incomparable. These works mainly focus on producing good margin guarantees; for similar guarantees for AdaBoost and related discussion, please the work of Telgarsky (2013).

Acknowledgements

This chapter is based on work by the dissertation author which appeared in the Journal of Machine Learning Research, Volume 13, pages 561{606, 2012.

Part II

Statistical Behavior

Chapter 4

The Primal and the Dual

Part II of this thesis considers statistical issues, namely controlling the difference between between L^b and L as a function of the sample size and various properties of the hypothesis class H, loss ℓ, and source distribution μ.

To warm up to this problem, this chapter merely provides a strong duality analog to the weak duality result Proposition 1.4.1. Unlike the strong duality result for finite samples (cf. Theorem 2.2.1), this duality result requires the loss ℓ to be Lipschitz, and consequently this section discusses some of the difficulties presented by the exponential loss.

4.1 Strong Duality

Without further ado, the strong duality result follows.

Lemma 4.1.1. Let $\ell : \mathbb{R} \to \mathbb{R}_+$ be convex with $\lim_{z \to 1} \ell(z) = 0$ and finite tightest Lipschitz constant $\sup_{x \neq y} |\ell(x) - \ell(y)|/|x - y| < 1$ be given. Additionally, let finite nonnegative measure μ over $X \times \{-1, +1\}$ and hypothesis class H be given. Then

$$\inf \left\{ \int \ell(A\lambda) d\mu : \lambda \geq 0 \right\} = \max \left\{ -\int \ell^*(-p) : p \in L^\infty(\mu); p \in [0, 1] \ \mu\text{-a.e.}, \|A^T p\|_\infty = 0 \right\}.$$

Additionally, $\lambda \geq 0$ is optimal if there exists $p \in L^\infty(\mu)$ such that $p(x,y) \in \partial\ell((A\lambda)_{x,y})$ for μ-a.e. (x,y) and p is feasible ($p \in [0, 1]$ μ-a.e., and $\|A^T p\|_\infty = 0$).

Comparing Lemma 4.1.1 to Theorem 2.2.1, the latter, in addition to handling non-Lipschitz losses, has an artifact term m. This in turn is a consequence of L^b using the "wrong" dual pairing $\langle p, f \rangle = \sum_i p_i f_i$ rather than $m^{-1} \sum_i p_i f_i$, a choice which was discussed in Remark 1.A.4.

In order to prove this duality result, the first step is to establish the convexity structure of L, which is analogous to the structure of L^b presented in Proposition 1.B.2, although once again L^b exhibits the artifact scaling term m.

Lemma 4.1.2. Let $\ell : \mathbb{R} \to \mathbb{R}_+$ be convex with $\lim_{z \to 1} \ell(z) = 0$ and finite tightest Lipschitz constant $\alpha := \sup_{x \neq y} |\ell(x) - \ell(y)| / |x - y| < 1$, and let μ be a probability measure over $Z := X \times \{-1, +1\}$.

1. If $f \in L^1(\mu)$, then $\int \ell(f(z)) d\mu(z)$ is well-defined and finite.
2. $\int \ell d\mu$ is convex lower semi-continuous over $L^1(\mu)$.
3. Its conjugate $(\int \ell d\mu)^*$ is also convex lower semi-continuous as a function over $L^1(\mu)$.
4. If $p \in L^1(\mu)$, then $(\int \ell d\mu)^*(p) = \int \ell^*(p) d\mu$, which is finite $(\int \ell d\mu)^*(p)$ if $p \in [0, \infty]$ μ-a.e..
5. If $f \in L^1(\mu)$, then $\partial(\int \ell d\mu)(f) \neq \emptyset$, and moreover $p \in \partial(\int \ell d\mu)(f)$ if $p(x) \in \partial\ell(f(z))$ for μ-a.e. z.

The proof of Lemma 4.1.2 can be expected to be somewhat similar to the analogous proofs for L^b in Proposition 1.B.2, however measurability issues prevent the integrals being treated identically to the finite sums in L^b. To circumvent measurability issues, the proof as usual must go through simple functions and limits thereof, which predictably adds some length.

From here, the proof of the strong duality rule in Lemma 4.1.1, proceeds by invoking an appropriately strengthened Fenchel duality rule, namely one strong enough to handle the Banach spaces in question (Zalinescu, 2002, Corollary 2.8.5 using condition (vii)). Note that this particular rule can not be applied if $\mathcal{H} = L^1(\mu)$ is replaced with, say, all weightings over H with finite support, which is not a complete normed vector space under the $L^1(\mu)$ norm.

Before turning to the complete proofs, the difficulties with the exponential loss stem from the fact that it is not in general finite over $L^1(\mu)$.

Proposition 4.1.3. Let μ denote the standard Gaussian measure over $\mathbb{R}$, and define $f(x) := x^2$ and $f_i(x) := x^2 \mathbf{1}[|x| \geq i]$. Then $f_i \in L^1(\mu)$, $f \in L^1(\mu)$, and $\|f_i - f\|_1 \to 0$, but

$$\int \exp(f_i(x)) d\mu(x) < 1 \qquad \text{and} \qquad \int \exp(f(x)) d\mu(x) = \infty.$$

Proof. To start, $\int f_i \, d\mu = 1$ (variance of a standard Gaussian), and thus $\int f_i \to f$ by the monotone convergence theorem (and so $\|f_i - f\|_1 \to 0$). But

$$\int \exp(f_i(x)) \, d\mu(x) \to \int e^{x^2} \, d\mu(x) < \infty;$$
$$\int \exp(f(x)) \, d\mu(x) = \frac{1}{\sqrt{2\pi}} \int e^{x^2/2} dx = \infty.$$

□

In particular, a step in the proof of Lemma 4.1.1 breaks down when this is combined with the possibility of A not being a closed subspace (cf. Lemma 1.A.6). This does not mean the situation with $\ell_{\exp}$ is hopeless; rather, it is simply the case that $L^1(\mu)$ (and dual space $L^\infty(\mu)$) is the wrong lens with which to inspect $\int \ell_{\exp} d\mu$.

Appendix 4.A Deferred Proofs

First, a proof of the convexity properties of $\int \ell \, d\mu$.

Proof of Lemma 4.1.2. Let $f \in L^1(\mu)$ be arbitrary. Since ℓ is convex and finite, it is continuous, so $\ell \circ f$ is measurable, and moreover it is nonnegative thus $\int \ell(f) d\mu$ is well-defined. Additionally,

$$\int \ell(f(z)) \, d\mu(z) = \int \ell(0) d\mu(z) + \int (\ell(f(z)) - \ell(0)) \, d\mu(z)$$
$$\ell(0)\,\mu(Z) + \int |f(z) - 0| \, d\mu(z) = \ell(0)\,\mu(Z) + \|f\|_1 < \infty.$$

Next, for any $f_1, f_2 \in L^1(\mu)$ and $\lambda \in [0,1]$,

$$\int \ell(\lambda f_1(z) + (1-\lambda) f_2(z)) \, d\mu(z) \leq \int (\lambda \ell(f_1(z)) + (1-\lambda) \ell(f_2(z))) \, d\mu(z)$$
$$= \lambda \int \ell(f_1(z)) \, d\mu(z) + (1-\lambda) \int \ell(f_2(z)) \, d\mu(z);$$

whereby $\int \ell \, d\mu$ is convex. Since it is finite over $L^1(\mu)$ (as above), it is necessarily lower semi-continuous.

Since $\int \ell$ is convex lower semi-continuous, so is its conjugate (Zalinescu, 2002, Theorem 2.3.3), where the dual space $L^1(\mu)$ is identified with $L^\infty(\mu)$ as per the isomorphism statements in Lemma 1.A.2.

The remainder of this proof will reason about the conjugate to $\int_R \grave{}\, d\,\cdot$. First let $p \in L^1(\,\cdot\,)$ be given with $(p^{-1}([0;\infty]^c)) > 0$; it will follow that $(\int_R \grave{}\, d\,\cdot)(p) = \infty$. Define the sets

$$S := p^{-1}((-\infty;0)) \qquad S_0 := p^{-1}([0;\infty]); \qquad S_+ p^{-1}((0;\infty));$$

as well as, for every $c \in \mathbb{R}$, the reals

$$g_- \in \partial\grave{}(-c); \qquad g_0 \in \partial\grave{}(0); \qquad g_+ \in \partial\grave{}(+c);$$

and lastly the simple functions

$$f_c(z) := -c\mathbf{1}[z \in S_-] + 0\mathbf{1}[z \in S_0] + c\mathbf{1}[z \in S_+];$$
$$g_c(z) := g_-\mathbf{1}[z \in S_-] + g_0\mathbf{1}[z \in S_0] + g_+\mathbf{1}[z \in S_+]:$$

By these choices f_c and g_c are measurable and within $L^1(\,\cdot\,)$, and moreover $g_c \in \partial\grave{}(f_c)$ everywhere. As such,

$$\begin{aligned}
\int \grave{}\, d\,(p) &= \sup\left\{\int (fp - \grave{}(f))\,d\,\cdot : f \in L^1(\,\cdot\,)\right\} \\
&\geq \sup\left\{\int (f_c p - \grave{}(f_c))\,d\,\cdot : c \in \mathbb{R}\right\} \\
&\geq \sup\left\{\int (f_c p - \grave{}(0) + g_c(0 - f_c))\,d\,\mu(z) : c \in \mathbb{R}\right\} \\
&\geq \sup\left\{c\left(\int_S (g_c - p)\,d\,\cdot + c\int_{S_+} (p - g_c)\,d\,\cdot\right) : c \in \mathbb{R}\right\} - \grave{}(0)\,\mu(Z) \\
&= \infty;
\end{aligned}$$

the last step following since $g_c \in [0;\infty]$ everywhere and $\mu(S \cup S_+) > 0$. As such, $(\int_R \grave{}\, d\,\cdot)(p) = \infty$, and since $\int_R \grave{}\,\mu(p)\,d\,\cdot = \infty$ by properties of $\grave{}$ (cf. Lemma 1.B.1), it follows that $\int_R \grave{}\,\mu(p)\,d\,\cdot = (\int_R \grave{}\, d\,\cdot)(p) = \infty$.

In the remainder of the proof of this conjugacy property, suppose $p \in [0;\infty]$ μ-a.e..

Now consider the case that p is a simple function with $p \in (0;\infty)$ everywhere. Since $\grave{}$ is finite over $[0;\infty]$ (cf. Lemma 1.B.1), p is within the relative interior of the domain of $\grave{}$ everywhere, and thus $\partial\grave{}(p(z))$ is a nonempty set for every $z \in Z$ (Rockafellar, 1970, Theorem 23.4).

Consequently, construct q so that $q(z) \in \partial \ell(p(z))$ everywhere, and moreover q is also a simple function (i.e., pick the same subgradient along each of the finitely many regions composing q); these choices will ensure that there are no measurability issues with q (otherwise, the arguments pass through for arbitrary $p \in (0, \infty)$); additionally, $q \in L^1(\mu)$ since μ is a finite measure. Since ℓ is lower semi-continuous, $q(z) \in \partial \ell(p(z))$ implies $p(z) \in \partial \ell^*(q(z)) = \partial(-q(z))$, and the Fenchel-Young inequality implies

$$\ell^*(p(z)) = p(z)q(z) - \ell(q(z)) = p(z)q(z) + \ell(-q(z)).$$

As such,

$$\begin{aligned} \int d\ell^*(p) &= \sup \left\{ \int (fp + \ell(-f))\, d\mu : f \in L^1(\mu) \right\} \\ &\geq \int (qp + \ell(-q))\, d\mu \\ &= \int \ell^*(p)\, d\mu \end{aligned}$$

Now using the fact that $p(z) \in \partial(-q(z))$,

$$\begin{aligned} \int d\ell^*(p) &= \sup \left\{ \int (fp + \ell(-f))\, d\mu : f \in L^1(\mu) \right\} \\ &\leq \sup \left\{ \int (fp + \ell(-q) - p(f-q))\, d\mu : f \in L^1(\mu) \right\} \\ &= \sup \left\{ \int (pq + \ell(-q))\, d\mu : f \in L^1(\mu) \right\} \\ &= \int \ell^*(p)\, d\mu; \end{aligned}$$

combining these two inequalities, $\left(\int d\ell^*\right)(p) = \int \ell^*(p)\, d\mu$.

Now consider the case that $p \in (0, \infty)$ is just measurable. Since the simple functions are dense in $L^1(\mu)$ (Folland, 1999, Theorem 6.8), there exists a simple function $\phi_i \in L^1(\mu)$ with $\|p - \phi_i\|_1 \leq 1/i$, and moreover ϕ_i may be clamped to the range $[1/i, 1/i]$ (with i sufficiently large to make this interval nonempty), whereby this clamped simple function ϕ_i satisfies $\|p - \phi_i\|_1 \leq 2/i$. Since $\int d\ell^*$ is lower semi-continuous,

$$\int d\ell^*(p) \leq \lim_i \int d\ell^*(\phi_i) = \lim_i \int \ell^*(\phi_i)\, d\mu = \int \ell^*(p)\, d\mu;$$

where the last step used the dominated convergence theorem applied with dominating constant map $z \mapsto \sup_{q \in [0, \|p\|]} |\ell(q)|$, which is finite since ℓ is continuous over the compact set $[0, \|p\|]$ (cf. Lemma 1.B.1).

Next consider the case that measurable $p \in (0, \infty)$ μ-a.e.; then $p(z) := p(z) 1[p(z) \in (0, \infty)] + (\infty) 1[p(z) \notin (0, \infty)]$ satisfies $(\int \ell \, d\mu)^*(p) = (\int \ell \, d\mu)^*(p)$ by definition of the conjugate (the integrals ignore measure zero sets), whereby $(\int \ell \, d\mu)^{**}(p) = (\int \ell \, d\mu)^*(p) = \int \ell^*(p) d\mu = \int \ell^{**}(p) d\mu$.

Lastly, suppose measurable $p \in [0, \infty]$ μ-a.e.. For each i, define $p_i = (1 - 1/i) p + \infty (2/i)$. Then $p_i \in (0, \infty)$ μ-a.e., and $\|p_i - p\|_1 \to 0$, whereby the lower semi-continuity of $(\int \ell \, d\mu)^{**}$ and dominated convergence theorem cover this case in the same way as the move away from simple functions.

Note lastly that these last choices provide a finite integral, since $\sup_{z \in [0, \|p\|]} |\ell(z)| < \infty$ as above, and μ is a finite measure.

Lastly, for the subdifferential rule, let f be given, and first suppose $p(z) \in \partial(f)$ for μ-a.e. z. Then, by the Fenchel-Young equality (applied to ℓ at each $f(z)$) and the conjugacy relation $(\int \ell)^* = \int \ell^*$ established above,

$$\int \ell(f(z)) \, d\mu(z) = \int (\ell^*(p(z)) + p(z) f(z)) \, d\mu(z) = \int \ell \, d\mu^*(p) + \int f p \, d\mu;$$

by the Fenchel-Young equality (now applied a single time to $\int \ell \, d\mu$ (Zalinescu, 2002, Theorem 2.4.2.iii)), this implies that $p \in \partial(\int \ell \, d\mu)(f)$. This establishes one direction of the desired iff, but also that the subdifferential is nonempty, since a valid choice of p always exists (for instance, take $p(z) \in \partial(f(z))$ when $|f(z)| < \infty$, and $p(z) = 0$ when $|f(z)| = \infty$, where the latter adjustment is made only over a null set since $f \in L^1(\mu)$).

The other direction is established via contrapositive; namely, let $p \in L^1(\mu)$ be given, and suppose the set

$$S := \{z \in Z : p(z) \notin \partial(f(z))\}$$

satisfies $\mu(S) > 0$. Since the Fenchel-Young inequality is an equality only in the case that a subdifferential relationship is satisfied, it follows that

$$\ell(f(z)) + \ell^*(p(z)) > p(z) f(z) \qquad \forall z \in S:$$

Consequently, since S has positive measure, and again using the relation $(\int_R \grave{} d\mu) = \int_R \grave{} \, d\mu$ as established above,

$$\begin{aligned}\int \grave{} \, d\mu(f) + \int \grave{} \, d\mu(p) &= \int_{S^c} \grave{}(f)\,d\mu + \int_{S^c} \grave{}\,(p)d\mu + \int_S \grave{}(f)\,d\mu + \int_S \grave{}\,(p)d\mu \\ &> \int_{S^c} fpd\mu + \int_S fpd\mu \\ &= \int_{S^c} fpd\mu ;\end{aligned}$$

thus, again invoking the fact that equality is only achieved in the Fenchel-Young inequality when a subdi erential relation is satis ed Zalinescu (2002, Theorem 2.4.2.iii), it follows that $p \in \partial(\int_R \grave{} \, d\mu)(f)$. □

Lastly, the proof of the duality relation is as follows.

Proof of Lemma 4.1.1. Consider the following two Fenchel problems:

$$\begin{aligned}V_p &:= \inf \int \grave{}(A\ )d\mu + \int 0\ d\mu\ :\ \in \ ; \\ V_d &:= \sup \int \grave{}\,(p)\ -\ _{\{0\}}(A^{>}p) : p \in L^1(\ )\ ;\end{aligned}$$

where $_{\{0\}}$ is the indicator for the set $\{0\}$,

$$_{\{0\}}(\ ) = \begin{cases} 0 & \text{when } = 0; \\ 1 & \text{otherwise,} \end{cases}$$

and is the conjugate to $\int_R 0\ d\mu$. In order to show $V_p = V_d$, also that attainment occurs in the dual, and lastly the optimality condition, an appropriate Fenchel duality rule will be applied (Zalinescu, 2002, Corollary 2.8.5 using condition (vii)), which requires the veri cation of the following properties.

First note that $\int_R \grave{}\,d\mu$ and $\int_R \grave{} \, d\mu$ are both convex lower semi-continuous, and moreover mutually conjugate (cf. Lemma 4.1.2). The function $\mapsto 0 = \int_R 0\,d\mu$ is immediately convex lower semi-continuous (over), and thus its conjugate $_{\{0\}}$ is similarly convex lower semi-continuous, and the two are mutually conjugate (Zalinescu, 2002, Theorem 2.3.3).

Both $L^1(\ )$ and $= L^1(\ )$ are Banach and therefore Frechet spaces. (The present proof is one of the reasons was taken to be a Banach space and not merely, say, weightings with nite support as used by the algorithm).

Let $\mathrm{dom}(f) = \{x : f(x) < 1\}$ denote the e ective domain of a convex function, meaning those values where it is nite. As provided by Lemma 4.1.2, $\mathrm{dom}(\int \grave{}\, d\ ) = L^1(\ )$, and thus, since $A : \ \to L^1(\ )$,

$$A(\mathrm{dom}(\ \ 7!\ 0)) \quad \mathrm{dom} \int \grave{}\, d \quad = A + L^1(\ ) = L^1(\ );$$

which settles the constraint quali cation. (Recall that A is not necessarily a closed subspace (cf. Lemma 1.A.6); thus problems would occur here if this proof were attempted for $\grave{} = \exp$, as $\mathrm{dom}(\int \grave{}\, d\ )$ would not swallow the closure issues of A.)

This completes the conditions necessary for the Fenchel duality result.

To adjust the statement into the desired form, Lemma 4.1.2 provided that $\int \grave{}\, d$ is nite i its input lies within $[0\ ;\ ]$ -a.e. (thus other values may safely be discarded from the optimization problem, which always has feasible point $0 \in L^1(\ )$), and secondly $_0(A^> p) < 1$ i $\|A^> p\|_1 = 0$ (recall the form of $\|A^> p\|_1$ in Lemma 1.A.5).

Lastly, for the optimality conditions, the invoked Fenchel duality rule provided by Zalinescu (2002, Corollary 2.8.5 using condition (vii)) also grants that $\ 2$ is optimal i there exists $p \in L^1(\ )$ which satis es $p \in @(\int \grave{}\, d\ )(A\ )$ and $A^> p \in @(\ \ 7! \int 0\, d\ )(\ )$. To translate this into the desired form, rst note that Lemma 4.1.2 provided $p \in @(\int \grave{}\, d\ )(A\ )$ i $p \in @\grave{}(A\ )$ -a.e.. Secondly, $@(\ \ 7! \int 0\, d\ ) = \{0\}$ everywhere (indeed, "everywhere" is the same as -a.e., where is counting measure over H and $= L^1(\ )$), thus $A^> p = 0$ everywhere, which is equivalent to $(A^> p)(\ ) = 0$ for all $\ 2$ by the form of $A^>$ developed in Lemma 1.A.5, which in turn is equivalent to $\|A^> p\| = 0$ again by Lemma 1.A.5. What remains is to necessitate $p \in [0;\ ]$ -a.e.. Since $\grave{}$ is Lipschitz with tightest constant , and $\grave{}$ is increasing, it follows that $@\grave{}\ \subseteq [0;\ ]$ everywhere, thus the above property $p \in @\grave{}(A\ )$ -a.e. grants one direction. On the other hand, if $p \notin [0;\ ]$ for a set of positive measure, then $p \notin @\grave{}(A\ )$ for a set of positive measure by Lemma 4.1.2, and thus the optimality condition $p \in @(\int \grave{}\, d\ )(A\ )$ can not hold. □

Appendix 4.B Bibliographic Remarks

While duality statements abound in machine learning, they appear to be rare when considering the optimization over the distribution. The results here were based on general Fenchel rules presented by Zalinescu (2002) and Borwein and Zhu (2005), however production of such general rules is the topic of great research in the convex analysis literature.

Acknowledgements

This chapter is based on work by the dissertation author which appeared in the Conference on Learning Theory, 2013.

Chapter 5

Finite Hypothesis Classes

This chapter provides the first set of statistical guarantees. In a sense, these results are strong, since they hold for a very relaxed class of convex losses | including not just $\ell_{\exp}$ and $\ell_{\log}$, but even ℓ_{russ}, which is not strictly convex, and ℓ_{hinge}, which is neither strictly convex nor differentiable | and moreover the algorithmic component is completely unspecified: the guarantees hold for solutions which are merely ϵ-optimal, for a provided value of ϵ.

What's the catch? The hypothesis class $\mathcal{H}$ must be finite. To put this in context, here are some ways to produce a boosting algorithm with statistical guarantees.

1. Add some sort of penalization or regularization to the (primal) objective function, as investigated by, for instance, Blanchard et al. (2003), Lugosi and Vayatis (2004).

2. Adjust the algorithm to take smaller steps, thus explicitly producing iterates with small norm as discussed in Chapter 1; this choice is investigated by Zhang and Yu (2005).

3. Apply an early stopping criterion, for instance the second option in Figure 1.3 (enforce $t \leq m^{\alpha}$); this option was investigated for the exponential loss by Bartlett and Traskin (2007), Schapire and Freund (2012), and is investigated in Chapter 6 for losses similar to the logistic loss.

4. This chapter says that it is enough to ensure $\mathcal{H}$ is finite, and that some black box minimization scheme is provided for the corresponding choice of $\mathcal{H}$ (and thus L), meaning the results of Chapter 3 suffice.

There is a productive way to work around this catch: start with an infinite class $\mathcal{H}_0$, and based on the availability of data, cut it down to some finite subset $\mathcal{H} \subseteq \mathcal{H}_0$. For many classes, it can be shown that as the sample size increases, the degradation incurred by this finitization goes to zero.

5.1 Overview

Throughout this chapter, will be identified with some R^n, where $n := |H| < 1$. Additionally, the relevant set of losses is as follows.

De nition 5.1.1. L_c contains all convex losses $\grave{} : R \to R_+$ which are positive at the origin, and satisfy $\lim_{z \to 1} \grave{}(z) = 0$.

This class contains all losses discussed in this thesis: $\grave{}_{exp}; \grave{}_{log}; \grave{}_{russ}; \grave{}_{hinge}$. Note that the algorithm in Figure 1.3 is only suited to the minimization of continuously differentiable choices, so for $\grave{}_{hinge}$ some other choice is needed, for instance subgradient descent, or some faster alternative (Nesterov, 2003).

As discussed previously, the guarantees will be roughly of the form: if a near-optimal solution over b is given, then the quality over will not be too bad.

Asking for a relationship between L^b and L, however, is in general too much to ask without further conditions on the provided solutions. In particular, Section 5.2 presents a negative result, namely that there exist nearly optimal choices over L^b which are moreover very stable (have good margin properties), however their behavior over L is arbitrary bad. This result determines the nature of the subsequent analysis and resulting bounds.

The solution is to once again rely on hard cores: the input space is broken into two pieces, one being the hard core, where there exist optimal choices that make a few mistakes, and the hard core's complement, where it is possible to have zero mistakes, albeit giving up on the existence of a minimizer to the true convex risk. This material appears in Section 5.3, and must make a few concessions as compared with the development in Chapter 2.

The hard core has direct entailments on the structure of the convex risk. Specifically, Section 5.4 establishes first that the true risk has quantifiable curvature over the hard core, and effectively zero error over the rest of the space; this mirrors the basic properties of hard cores for finite samples provided by Theorem 2.4.8. This section then goes on to show how, with high probability, this structure carries over to any sampled instance.

The significance of first proving properties of the true risk, and then carrying them over to the sample, is that quantities dictating the structure of the empirical convex risk are sample independent Consequently, finite sample guarantees, which appear in Section 5.5, display a number of terms which are properties of the true convex risk, and not simply opaque random

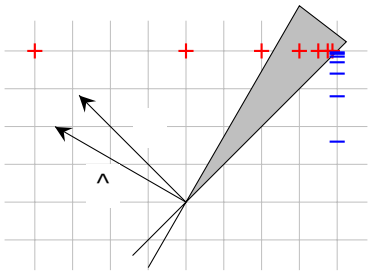

Figure 5.1. Unconstrained minimization can have some bad statistical properties; please see Proposition 5.2.1 and the surrounding discussion.

variables derived from the sample.

5.2 An Impossibility Result

The stated goal of allowing iterates to have unbounded norms is at odds with the task of bounding the convex risk L.

Proposition 5.2.1. There exists a probability measure , input space X, and nite set of hypotheses H with the following characteristics.

1. X is the square $[-1;+1]^2$, and H consists of the two projection maps.
2. has countable support.
3. There exists a perfect separator, albeit with zero margin.
4. For any $\ell \in L_c$, $L(A\ell) = 0$.
5. Let any nite sample $f(x_i;y_i)g_{i=1}^m$, any $b > 0$, and any $\ell \in L_c$ be given. Then there exists a maximum margin solution $\hat{}$, i.e., a solution satisfying

$$\arg\min_{i\in[m]} \frac{y_i(H\hat{})(x_i)}{k\hat{}k_1} = \sup\left(\arg\min_{i\in[m]} y_i(H)(x_i) : \in R^n; k k_1 = 1 \right);$$

which has $R^b(H\hat{}) = 0$ and $L^b(A) \leq 1=b$ but $L(A\hat{}) \geq b$.

A full proof is provided in Section 5.C, but the mechanism is simple enough to appear as a picture. Consider the linear classi cation problem in Figure 5.1, which has positive ("+") and negative ("-") examples along two lines. Optimal solutions to R are of the form c, where $= (-1;+1)$ and $c > 0$ (note $\lim_{c\to\infty} L(cA) = L(A) = 0$). Unfortunately, the positive and

negative examples are staggered; as a result, for any sample, every max margin predictor $\hat{w}$ which is determined solely by the rightmost "+" and uppermost "-", will fail to agree with the optimal predictor on some small region. A positive probability mass of points fall within this region, and so, by considering scalings $c\hat{w}$ as $c \to \infty$, the convex risk $\mathcal{L}$ may be made arbitrary large.

The statement of Proposition 5.2.1 is encumbered with details in order to convey the message that not only do such examples exist, they are fairly benign; indeed, the example depends on the additional regularity of large margin solutions. The only difficulty is the lack of any norm constraint on permissible iterates.

On the other hand, notice that the classification risk R is not only small, but its empirical counterpart $\hat{R}$ provides a reasonable estimate as n increases. Furthermore, if the distribution were adjusted slightly so that every $w \in \mathbb{R}^n$ made some mistake, then these unbounded iterates would fail to exist: the huge penalty for predictions very far from correct would constrain the norms of all good predictors.

The preceding paragraph describes the exact strategy of the remainder of this chapter: linear classification problems are split into two pieces, one where optimization may produced unboundedly large iterates with small classification risk, and another piece where iterates are bounded thanks to the presence of difficult examples.

5.3 Hard Cores

The choice here is to split the problem along a hard core. Recall, from Section 1.7, that the hard core was the support $[p > 0]$ of a reweighting $p \in L^1(\mathcal{D})$ which decorrelated every predictor, meaning $\int_R (A \circ \pi) p \, d\mathcal{D} = 0$. The choice of $L^1(\mathcal{D})$ matches the strong duality result for Lipschitz losses in Lemma 4.1.1; while this may seem to conflict with the generality of L_c, which contains non-Lipschitz losses, strong duality will only be invoked in proofs of hard core properties, and need only apply for a single loss. Thereafter, when applying these properties, the finiteness of $\mathcal{H}$ will be essential. Recall additionally from Section 1.7 that hard cores are guaranteed to exist (cf. Theorem 1.7.4).

The following result shows that hard cores achieve the goal laid out at the closing of Section 5.2; this statement can be seen as a distillation of the components of Theorem 2.4.8 which are needed for the statistical proofs here. The proof of this statement, which is somewhat involved, appears in Section 5.D.

Theorem 5.3.1. Let nite hypothesis class H, probability measure , and hard core C be given. The following statements hold.

1. There exists a sequence $\{h_i\}_{i=1}^{1}$ with $y(H\ h_i)(x) = 0$ for -a.e. $(x; y) \ 2\ C$, and $y^0(H\ h_i)(x^0) \ "$ 1 for -a.e. $(x^0; y^0) \ 2\ C^c$.

2. Every $\ 2\ R^n$ satis es either $(C \setminus [y(H\ )(x) = 0]) = \ (C)$ or $(C \setminus [y(H\ )(x) < 0]) > 0$.

The rst property provides the existence of a sequence which is not only very good -a.e. over C^c, but furthermore does not impact the value of H over C; that is to say, this sequence can grow unboundedly, and have unboundedly positive margins over C^c, while optimization over C can e ectively proceed independently. On the other hand, C is di cult: every predictor is either abstaining -a.e., or makes errors on a set of positive measure.

Finally, corresponding to the hard core, it will be convenient in future sections to use adopt the following notation for risk specialized to some region.

De nition 5.3.2. Given a set C (typically C or C^c), loss $`$, function class F $\ L^1(\ )$, and any $f \ 2\ F$, de ne

$$L(f\ ; C) := \int_C `(f(x; y))\, d\ (x; y) \qquad \text{and} \qquad L(F\ ; C) := \inf_{f\ 2F} L(f\ ; C);$$

with analogous de nitions for L^b, and additionally for R and R^b over functions in $L^1(\ ^X)$.

5.4 Hard Cores and Convex Risk

The hard core imposes the following structure on L. As provided by Theorem 5.3.1, there is a sequence which does arbitrarily well over C^c, without impacting predictions over C. On the other hand, since mistakes must occur over C, convex losses within L_c will be forced to avoid large predictors.

Theorem 5.4.1. Let nite hypothesis class H, probability measure , hard core C, and loss $` \ 2\ L_c$ be given.

1. There exists a sequence $\{h_i\}_{i=1}^{1}$ with $y(H\ h_i)(x) = 0$ for -a.e. $(x; y) \ 2\ C$, and moreover $\lim_{i\ !1} \ `((A\ h_i)_{x^0;y^0}) = 0$ for -a.e. $(x^0; y^0) \ 2\ C^c$.

2. Let any $\epsilon > 0$ be given. Then there exists $c \in \mathbb{R}$ and a set N with $\mu(N) = 0$ so that for every $\gamma \in \mathbb{R}^n$ with $L(A(\gamma); C) \leq L(A(\gamma'); C) + \epsilon$, there exists a representation $\gamma' \in \mathbb{R}^n$ with $H_\gamma = H_{\gamma'}$ over $C \setminus N$, and $\|\gamma'\|_1 \leq c$.

The structural properties of the true convex risk transfer over, with high probability, to any sampled problem. Crucially, the various bounds are quantified outside the probability; that is to say, they do not depend on the sample.

Theorem 5.4.2. Let finite hypothesis class H, probability measure μ, hard core C, and loss $\ell \in L_c$ be given.

1. With probability 1 over the draw of a finite sample, there exists $\gamma \in \mathbb{R}^n$ so that every $(x_i, y_i) \in C^c$ satisfies $y_i(H_\gamma)(x_i) > 0$, and every $(x_i', y_i') \in C$ satisfies $y_i'(H_\gamma)(x_i') = 0$.

2. Given any empirical suboptimality $\epsilon > 0$, there exist $c > 0$ and $b > 0$ so that for any $\delta > 0$, with probability at least $1 - \delta$ over a draw of m points where m_C, the number of points landing in C, has bound

$$m_C \leq c^2(\ln(n) + \ln(1/\delta));$$

then every ϵ-suboptimal $\gamma \in \mathbb{R}^n$ over the sample restricted to C, meaning

$$\hat{L}_b(A(\gamma); C) \leq \hat{L}_b(A(\gamma'); C) + \epsilon;$$

has a representation γ' with $\|\gamma'\|_1 \leq b$ which has $H_\gamma = H_{\gamma'}$ over the sample restricted to C, and in general μ-a.e. over C.

5.5 Deviation Inequalities

With the structure of the convex risk in place, the stage is set to establish deviation inequalities. These will be stated in terms of both a convex risk L, but also the classification risk R. In order to make this correspondence, this chapter relies on the following standard techniques due to Zhang (2004) and Bartlett, Jordan, and McAuliffe (2006).

Definition 5.5.1. Let $\mathcal{F}$ denote the set of measurable functions over X, and let $\mathcal{F}_Y$ denote the family of maps $f(x, y) \mapsto yf(x) : f \in \mathcal{F}\}$. (Note that H is to A as $\mathcal{F}$ is to $\mathcal{F}_Y$: this definitions exist for convenient use of L.)

Proposition 5.5.2. (See also Bartlett et al. (2006).) Let any $\ell \in L_c$ be given with ℓ differentiable at 0. There exists an associated function $\psi : [0,1] \to [0,1]$ with the following properties. First, for any probability measure μ and any $f : X \to R$, $(R(f) - R(F)) \le (x,y) \mapsto yf(x)) \le (F_Y)$. Second, the inverse ψ^{-1} exists over $[0,1)$, and satisfies $\psi^{-1}(r) \to 0$ as $r \to 0$.

Definition 5.5.3. Given $\ell \in L_c$ with ℓ differentiable at 0, let ψ, called the ψ-transform, be as in Proposition 5.5.2.

The general use of ψ is through its inverse, which provides, for any ℓ satisfying the conditions to Proposition 5.5.2,

$$\begin{aligned} R(H) - R(F) &\le \psi^{-1}(L(A) - L(F_Y)) \\ &= \psi^{-1}(L(A) - L(A) + L(A) - L(F_Y)). \end{aligned}$$

Note firstly that ψ^{-1} may be unwieldy, it is frequently easy to provide a useful upper bound. For instance, the exponential loss has $\psi^{-1}(r) \le 2\sqrt{r}$, the logistic loss has $\psi^{-1}(r) \le 4\sqrt{r}$, and the hinge loss has $\psi^{-1}(r) = r$ (Zhang, 2004, Bartlett et al., 2006). Secondly, the value of the second decomposition above is that by choosing H large enough, $L(A) - L(F_Y)$ may be forced arbitrarily small.

Theorem 5.5.4. Let finite hypothesis class H, probability measure μ, hard core C, and loss $\ell \in L_c$ be given. Let a suboptimality tolerance $\epsilon > 0$ be given; results will depend on reals $c > 0$ and $b > 0$ determined by the preceding terms. The following statements simultaneously hold with any probability $1 - \delta$ over the draw of m samples (with $\delta_0 := \delta/8$ for convenience), and any weighting $\alpha \in R^n$ which is ϵ-suboptimal (with $\epsilon \le$) for the corresponding surrogate empirical risk problem, meaning $\hat{L}_b(A) \le \hat{L}_b(A) + \epsilon$.

1. Let m_C and m_+ respectively denote the number of samples falling into C and C^c. Then

$$\begin{aligned} m_C &\le m\mu(C) + \sqrt{\ln(1/\delta_0)/(2m)}, \\ m_+ &\le m\mu(C^c) + \sqrt{\ln(1/\delta_0)/(2m)}. \end{aligned}$$

2. The true classification risk over the unbounded portion, C^c, has bound

$$\frac{R(H\ ;C^c)}{(C^c)}\quad \frac{}{(0)}+2\ \sqrt{\frac{2\ (n\ln(2m_+ +1)+\ln(4\ =\ ^0))}{(0)m_+}}+\frac{4(n\ln(2m_+ +1)+\ln(4\ =\ ^0))}{m_+}: \tag{5.5.5}$$

If moreover $<\ `\ (0)=m$, then

$$\frac{R(H\ ;C^c)}{(C^c)}\quad \frac{4(n\ln(2m_+ +1)+\ln(4\ =\ ^0))}{m_+}: \tag{5.5.6}$$

3. Suppose

$$m_C\quad c^2(\ln(\ n)+\ln(6\ =\ ^0)):$$

The true surrogate risk over the bounded portion, C, has bound

$$L(A\ ;C)\ \ L\ (A\ ;\ C)\quad +\frac{c\ \sqrt{\ln(n)}+4\sqrt{\ln(2=\ ^0)}}{\sqrt{m_C}}; \tag{5.5.7}$$

Additionally, if $`$ is differentiable at 0, the classification risk has bound

$$R(H\ ;C)\ \ R\ (F_Y;C)\quad {}^1\ +\left(\frac{c\ \sqrt{\ln(n)}+4\sqrt{\ln(2=\ ^0)}}{\sqrt{m_C}}\right)^!$$
$$+\ L(A\ ;\ C)\ \ L\ (F_Y;C)\ : \tag{5.5.8}$$

4. Suppose, for simplicity, that

$$m\quad \max\ \ 2\ln(1=\ ^0)=\min\{\ (C)^2;\ \ (C^c)^2g;2c^2(\ln(\ n)+\ln(1\ =\ ^0))=\ (C)$$

(where bounds are interpreted to hold trivially when denominators contain 0) and additionally that $<\ `\ (0)=m$ and $`$ is differentiable at 0. Then the true classification risk of the full

problem has bound

$$R(H) - R(F) \leq \left(1 + \frac{c\sqrt{2}\sqrt{\ln(n)} + 4\sqrt{\ln(2/\delta)}}{\sqrt{m\,\Phi(C)}}\right) \cdot L(\mathrm{span}(H), C) - L(F_Y, C) + \frac{8(n \ln(m\,\Phi(C^c)) + 1) + \ln(4/\delta)}{m\,\Phi(C^c)}.$$

As a consequence of these bounds, it is reasonable to choose H so that $\ln(|H|) = o(m)$; for instance, $|H| \approx \exp(\sqrt{m})$ gives roughly a rate of $m^{-1/4}$.

Appendix 5.A Technical Preliminaries

Lemma 5.A.1. Let any $\ell \in L_c$ be given. Then ℓ is continuous, measurable, and nondecreasing. Subgradients exist everywhere, and satisfy $\partial\ell(0) \subseteq \mathbb{R}_{++}$.

Proof. Since ℓ is finite everywhere, it is continuous (Rockafellar, 1970, Corollary 10.1.1), and thus measurable (Folland, 1999, Corollary 2.2). Since convex functions are subdifferentiable everywhere along the relative interior of their domains (which in this case is just $\mathbb{R}$), it follows that ℓ has subgradients everywhere (Rockafellar, 1970, Theorem 23.4).

If ℓ were not nondecreasing, there would exist $x < y$ with $\ell(x) > \ell(y)$; but that means every subgradient $g \in \partial\ell(x)$ satisfies

$$\ell(y) \geq \ell(x) + g(y - x),$$

and thus $g < 0$. But then, for any $z < x$, $\ell(z) \geq \ell(x) + g(z - x)$, which in particular contradicts $\lim_{z \to -\infty} \ell(z) = 0$ (indeed, it implies $\lim_{z \to -\infty} \ell(z) = -\infty$), thus ℓ is nondecreasing.

Next, since ℓ is nondecreasing, $\partial\ell \subseteq \mathbb{R}_+$. However, since $\ell(0) > 0$, it follows that $\partial\ell(0) \subseteq \mathbb{R}_{++}$, since otherwise $\lim_{z \to -\infty} \ell(z) = 0$ would be contradicted. □

The following grants continuity of the map $\ell \mapsto \int \ell(A)\,d\mu$ for all of L_c, and in this way is stronger than the analogous result for general choices of H (cf. Lemma 4.1.2). Unsurprisingly,

the proof crucially hinges upon $= R^n$ being nite dimensional (in particular, the compactness of l_1 balls is used, which is of course in general false in the l^1 topology in general).

Lemma 5.A.2. Let nite hypothesis class H, nite measure , and loss $` 2 L_c$ be given. Then the map $7! \int `(A)d$ is convex and continuous.

Proof. Convexity follows from convexity of $`$; namely, let ; ^{0}, and $2 [0;1]$ be given; then

$$\begin{aligned}\int `( y(H( + (1 ) ^0))_x)d (x;y) &= \int `(y (H)_x (1)y(H ^0)_x)d (x;y)\\ &\int (` ( y(H )_x) + (1 )`(y(H ^0)_x)) d (x;y)\\ &= \int `( y(H )_x)d (x;y)\\ &\quad + (1 ) \int `(y(H ^0)_x)d (x;y):\end{aligned}$$

For continuity, if su ces to show that the map commutes with limits. In particular, let a convergent sequence $_i !$ be given, and without loss of generality suppose there exists (nite) $B 0$ so that $k _i k_1 B$ (whereby $k k_1 B$ since $k _i k_1 ! 0$). Since $\sup_{h;x;y} jyh(x)j 1$ and $`$ is nondecreasing, it follows for every i and $(x;y) 2 X f 1;+1g$ that

$$`( y(H _i)_x) `(k _i k_1) `(B):$$

Furthermore, note that the map $f (x;y) = `(B)$ satis es $f 2 L^1( )$ since is a nite measure, i.e., $kf k_1 `(B) (X f 1;+1g) < 1$. Consequently, by the dominated convergence theorem with dominating map f and the continuity of $`$ (cf. Lemma 5.A.1),

$$\lim_{i!1} \int `( y(H _i)_x)d (x) = \int \lim_{i!1} `(y(H _i)_x)d (x) = \int `(y(H)_x)d (x)$$

as desired. □

Proposition 5.A.3. Let nite hypothesis class H, probability measure , and loss $` 2 L_c$ be given. Then given a bound $b$ on the $l_1$ norm of considered predictors, there exists $c `(b)$ so that, for any > 0, with probability at least 1 over the draw of m points from , every $2 R^n$ with $k k_1 b$ satis es

$$L(A) \hat{L}_b(A) \frac{c^{\sqrt{\ln(n)}} + \sqrt{\ln(2=)}}{\sqrt{m}}:$$

Proof. Let bound b and loss $\ell \in L_c$ be given. Define a truncation

$$\hat{\ell}(z) := \begin{cases} \ell(z) & \text{when } z \leq b, \\ \ell(b) & \text{otherwise.} \end{cases}$$

Since ℓ is nondecreasing (cf. Lemma 5.A.1), $\hat{\ell}(z) \leq \ell(b)$, and furthermore $\hat{\ell}$ is Lipschitz with a constant that may be measured at b; indeed, since ℓ is finite everywhere, it has bounded subdifferential sets (Rockafellar, 1970, Theorem 23.4), and thus, taking any $z_1, z_2 \in \mathbb{R}$ and supposing without loss of generality that $z_1 \leq z_2$,

$$\begin{aligned} |\hat{\ell}(z_2) - \hat{\ell}(z_1)| &= \hat{\ell}(z_2) - \hat{\ell}(z_1) \\ &\leq \sup\{\ell(z_2) - (\ell(z_2) + \langle g_2, z_1 - z_2 \rangle) : g_2 \in \partial(z_2)\} \\ &= |z_2 - z_1| \sup\{|g_2| : g_2 \in \partial(z_2)\} \\ &< \ell' |z_2 - z_1|, \end{aligned}$$

correspondingly, set a Lipschitz constant $L_{\hat{\ell}} := \sup\{|g| : g \in \partial(b)\}$.

Note that for every $\alpha \in \mathbb{R}^n$ with $\|\alpha\|_1 \leq b$, $\sup_{x \in X} |(H\alpha)_x| \leq b$, and thus $L(A_\alpha) = \int \hat{\ell}(A_\alpha) d$ and $\hat{L}_b(A_\alpha) = \int \hat{\ell}(A_\alpha) db$. Lastly, the desired constant c, which does not depend on ℓ, n, or m, will be $c := \max\{2L_{\hat{\ell}} b \sqrt{2}, \ell(b)\}$.

Now let a sample of size m be given, and let let $R_m(\mathrm{span}(H); b)$ denote the Rademacher complexity of $\{H\alpha : \|\alpha\|_1 \leq b\}$. By properties of Rademacher complexity and a few appeals to McDiarmid's inequality, (Boucheron et al., 2005a, Theorem 3.1, and the proof of Theorem 4.1), with probability at least $1 - \delta$ over the draw of this sample,

$$\begin{aligned} \sup_{\|\alpha\|_1 \leq b} L(A_\alpha) - \hat{L}_b(A_\alpha) &= \sup_{\|\alpha\|_1 \leq b} \int \hat{\ell}(A_\alpha) d - \int \hat{\ell}(A_\alpha) db \\ &\leq 2L_{\hat{\ell}} R_m(\mathrm{span}(H); b) + \sqrt{\frac{2\ln(2/\delta)}{m}}. \end{aligned} \tag{5.A.4}$$

Next, taking $R_m(S)$ denote the Rademacher complexity of a set of functions S, by $R_m(\mathrm{span}(H); b) = bR_m(\mathrm{span}(H); 1) = bR_m(H)$ and an appeal to Massart's Finite Lemma (Boucheron et al., 2005a, Theorem 3.3)

$$R_m(\mathrm{span}(H); b) \leq b\sqrt{\frac{2\ln(n)}{m}}.$$

Plugging this into eq. (5.A.4) and recalling the choice $c = \max\{2L\sqrt{b^{p}\bar{2}}, \ell(b)\}$, the result follows. □

Lemma 5.A.5. Let $S \subseteq R$ and convex $f : S \to R$ be given. If $x, y \in S$ are given with $x < y$ and $f(x) < f(y)$, then for every $S \ni z > y$, $f(y) < f(z)$.

Proof. Write y as a combination of x and z:

$$y = x\,\frac{z-y}{z-x} + z\,\frac{y-x}{z-x}.$$

By convexity and $f(y) > f(x)$,

$$f(y)\,\frac{z-y}{z-x} + f(z)\,\frac{y-x}{z-x} > f(x)\,\frac{z-y}{z-x} + f(z)\,\frac{y-x}{z-x}$$
$$f\left(x\,\frac{z-y}{z-x} + z\,\frac{y-x}{z-x}\right)$$
$$= f(y).$$

Rearranging and using $x < y$, it follows that $f(y) < f(z)$. □

Appendix 5.B Structure of L over $S_D(H, \cdot)$

The following theorem leads to a number of properties presented in Sections 5.3 and 5.4; it is easiest to prove them at once, as a ring of implications. Note the resemblance to Theorem 2.4.4 from Chapter 2.

Theorem 5.B.1. Let finite hypothesis class H, probability measure $\cdot$, and a set D be given. The following statements are equivalent.

1. For every $\in R^{n}$, either $(D \setminus [y(H\ )x = 0]) = (D)$ or $(D \setminus [y(H\ )x < 0]) > 0$.
2. Given any and any $\ell \in L_c$, there exists a bound b and a null set $N \subseteq X$ $\{-1, +1\}$ (i.e., $(N) = 0$) so that for every -suboptimal weighting $\hat{}$ over D, meaning any weighting satisfying

$$L(A\ ; D) \quad L\ (A\ ; D) + \ ;$$

there exists ^{0} with $k \quad k_1 \quad b$ and $H^{\hat{}} = H^{\ 0}$ over $D \cap N$.

3. $D \in S_D(H, \cdot)$.

The following structural lemma is crucial. The finiteness of H is used in two places. Less essentially, it ensures the existence of orthogonal projections (since subspaces of $\mathbb{R}^n$ are always closed), but this could be circumvented with points that are almost closest. The more crucial reliance is in exhibiting a positive constant $c > 0$ on the difference of norms between H and $\cdot$ (over $K^\perp$); this in turn depends on the compactness of the L_1 ball, which is false for general vector spaces (with compactness measured in the L^1 topology).

Lemma 5.B.2. Let finite hypothesis class H, probability measure $\cdot$, and a set D be given. Define the set

$$K := \{ \in \mathbb{R}^n : y(h - h')x = 0 \text{ for } \cdot\text{-a.e. } (x, y) \in D\}.$$

The following statements hold.

1. K is a closed subspace, as is its orthogonal complement $K^\perp$, and orthogonal projection onto $K^\perp$ is well-defined.
2. There exists a set N with $(N) = 0$ so that, for any for any $\in \mathbb{R}^n$, the orthogonal projection $\mapsto {}^\perp \in K^\perp$ satisfies $h - h' = h - h'^\perp$ everywhere over $D \setminus N$.
3. There exists a constant $c > 0$ so that, for any $\in \mathbb{R}^n$ with $(D \setminus [A \neq 0]) > 0$, $\|A\|_{L^1(\cdot_D)} / \|^\perp\|_1 > c$, where $L^1(\cdot_D)$ is the L^1 metric with respect to the measure defined by $\cdot_D(S) = (D \setminus S)$ for any measurable set S.

Proof. (Item 1) Direct from its construction, K is a subspace. Since $\mathbb{R}^n$ is finite dimensional, it follows that both K and $K^\perp$ are closed subspaces, whereby the orthogonal projection $\mapsto {}^\perp$ exists.

(Item 2) Given the subspace pair K and $K^\perp$, for any $\in \mathbb{R}^n$, there exists the decomposition $\mapsto {}^K + {}^\perp$, where ${}^\perp \in K^\perp$. By definition, $h - h'^K = 0$ $\cdot$-a.e. over D, and thus $h - h' = h - h'^\perp$ $\cdot$-a.e. over D.

Now let Q be any countable dense subset of $\mathbb{R}^n$. For each $_i \in Q$, define $N_i := [h - h'_i \neq h - h'^\perp_i]$, where the above provides $(N_i) = 0$. Set $N := \bigcup_i N_i$, which is measurable since it is a countable union, and moreover $(N) = 0$ by $\cdot$-additivity. It will now be argued that the projections onto $K^\perp$ give equivalences over $D \setminus N$.

To this end, let any $\in R^n$, any $(x;y) \in D \setminus N$, and any > 0 be given. Since Q is a countable dense subset of R^n, there exists $_i \in Q$ with $k_i \quad k_2 \quad k_i \quad k_1 \quad =2$. Now let $P^?$ denote the orthogonal projection operator onto $K^?$; by the triangle inequality, Cauchy-Schwarz, and since orthogonal projection is a contraction,

$$\begin{aligned} 0 \quad j(H \quad)(x) \quad (H \quad ^?)(x)j &= j(H \quad)(x) \quad (HP^? \quad)(x)j \\ &= j(H(\quad _i + \quad _i)(x) \quad (HP^?(\quad _i + \quad _i))(x)j \\ &\quad j(H \quad _i)(x) \quad (H \quad ^?_i)(x)j + jH(\quad _i)(x)j + jHP^?(\quad _i)(x)j \\ &\quad j \quad 0j + \sup_{h;x} jh(x)jk \quad _i k_1 + \sup_{h;x} jh(x)jkP^? k_2 k \quad _i k_2 \\ &\quad 0 + \quad =2 + \quad =2 = \quad : \end{aligned}$$

Taking $\# 0$, it follows that $H \quad = H \quad ^?$ over $D \setminus N$.

(Item 3) For the final part, if every $\in R^n$ has $(D \setminus [A \not\equiv 0]) = 0$, there is nothing to show, so suppose there exists $\in R^n$ with $_D([A \not\equiv 0]) > 0$. Consider the optimization problem

$$\inf \quad \frac{kA \quad k_{L^1(\ _D)}}{k \quad ^? k_1} : \quad \in R^n; \ _D([A \not\equiv 0]) > 0 \quad = \inf \quad kA \quad k_{L^1(\ _D)} : \quad \in K^?; k \quad k_1 = 1 \quad :$$

The latter is a minimization of a continuous function over a nonempty compact set, and thus attains a minimizer $$. But $\in K^?$ and $k \quad k_1 = 1$, thus $kA \quad k_{L^1(\ _D)} > 0$. The result follows with $c := kA \quad k_{L^1(\ _D)} > 0$. □

Proof of Theorem 5.B.1. (Item 1 =) Item 2.) The proof will proceed by contradiction. Specifically, suppose Item 1, but contradictorily that Item 2 fails to hold; in particular, there exists $\in R$ and $` \in L_c$ so that for any $b \quad 0$ and any null set $N \quad X \quad f \quad 1;+1g$, there exists $\hat{} \in R^n$ with $L(A\hat{}) \quad L(A; D) + \ $, but there is no representation $^0 \in R^n$ satisfying $k \ ^0 k_1 \quad b$ and $H\hat{} = H \quad ^0$ over $D \setminus N$. As such, let N be the null set, as provided by Lemma 5.B.2, so that every $\in R^n$ has $H \quad = H \quad ^?$ everywhere on $D \setminus N$. Additionally, let $$ and $`$ be as given by the contradiction; therefore, for every $i \in Z_{++}$, there exists $_i$ with $L(A \ _i; D) \quad L(A; D) + \ $, but there is no representation ^0_i with $H \quad _i = H \quad ^0_i$ over $D \setminus N$ and $k \ ^0_i k_1 \quad i$.

For every i, by Lemma 5.B.2 and the choice of N, the projection $^?_i$ of $_i$ onto $K^?$ satisfies $H \quad _i = H \quad ^?_i$ over $D \setminus N$ (whereby $L(A \ ^?_i; D) \quad L(A; D) + \ $ still holds), and thus the

above assumed contradiction provides $k \, ^{?}_{i} k \quad i$. Note that $f \, ^{?}_{i} = k \, ^{?}_{i} k_1 g_{i=1}^{1}$ lies in a compact set (the unit l^1 ball), and thus let $f \, ^{(2)}_{i} g_{i=1}^{1}$ be a subsequence of $f \, ^{?}_{i} g_{i=1}^{1}$ with $^{(2)}_{i} = k \, ^{(2)}_{i} k_1 \, ! \quad 2 R^n$. Since $K^{?}$ is a closed subspace, it follows that $2 K^{?}$ as well, and furthermore $k \, k_1 = 1$. Since $k \, ^{?}_{i} k_1 \quad i$ and the subsequence of $^{(2)}_{i} g_{i=1}^{1}$ can be formed simply via deletions, it follows that $k \, ^{(2)}_{i} k_1 \quad i$, and in particular $\lim_{i!1} \quad k \, ^{(2)}_{i} k_1 = 1$, and lastly $L(A \, ^{(2)}_{i}; D) \quad L \, (A \, ; \, D) +$ still holds. Furthermore, since Lemma 5.B.2 also provides a $c > 0$ with $kA \, k_{L^1(\ D)} = k \, k_1 \quad c$ for $2 K^{?}$, it follows that $kA \, k_{L^1(\ D)} \quad c$, which is only possible if $(D \setminus [y(H \ )x \ 0]) > 0$.

By assumption (i.e., by Item 1), since $(D \setminus [y(H \ )x \ 0]) > 0$, then $(D \setminus [y(H \ )x < 0]) > 0$; for convenience, define the set $P := [y(H \ )(x) < 0]$. Thus, for any $2 R^n$, taking any $g 2 @`(0)$ (note $g > 0$ via Lemma 5.A.1),

$$
\begin{aligned}
&\lim_{t!1} \frac{\int_D `( \ y(H( \ + t \ ))(x))\, d \ (x;y) \quad \int_D `(\ y(H \ )(x))\, d \ (x;y)}{t} \\
&\quad \lim_{t!1} \frac{\int_{D \setminus P} `( \ y(H( \ + t \ ))(x))\, d \ (x;y) \quad \int_D `(\ y(H \ )(x))\, d \ (x;y)}{t} \\
&\quad \lim_{t!1} \frac{\int_{D \setminus P} (`(0) + g( \ y(H( \ + t \ ))(x)))\, d \ (x;y) \quad \int_D `(\ y(H \ )(x))\, d \ (x;y)}{t} \\
&= g \int_{D \setminus P} y(H \ )(x) d \ (x;y) \\
&\quad + \lim_{t!1} \frac{\int_{D \setminus P} (`(0) + g( \ y(H \ )(x)))\, d \ (x;y) \quad \int_D `(\ y(H \ )(x))\, d \ (x;y)}{t} \\
&> 0: \qquad (5.B.3)
\end{aligned}
$$

The above derivation shows that $\int_D `(A(\ + t \ ))\, d$ grows arbitrarily large with t, and in particular the ray $f \ + t \ : t \quad 0g$ must exit (and never return to) the desired $$-sublevel set

$$C := f \quad 2 R^n : L(A \ ; D) \quad L \, (A \, ; \, D) + \ g:$$

To develop the contradiction, it will be shown that the construction of $$ indicates that this ray should in fact not be leaving C; the proof will be similar to one due to Hiriart-Urruty and Lemaréchal (2001, Proposition A.2.2.3).

Since $\int_D `d$ is convex and continuous (via Lemma 5.A.2, using finite measure $_D$ defines as $_D(S) = \ (D \setminus S)$ for all measurable $S$), the sublevel sets of $\int_D `d$ are closed convex sets, and in particular C is closed and convex. By construction of $$ and continuity of H, recalling $\lim_{i!1} \quad k \, ^{(2)}_{i} k_1 = 1$ and $^{(2)}_{i} \ 2 \ C$, and using the closed convexity of C, every $^{0} 2 \ C$ and

$t > 0$ satisfies

$$H^0 + tH = \lim_{i \to \infty} \left(1 - \frac{t}{k_i^{(2)} k_1}\right) H^0 + \frac{t}{k_i^{(2)} k_1} H_i^{(2)} \in C :$$

This holds for all $t > 0$, which contradicts the fact, derived in eq. (5.B.3), that the ray $\{H^0 + t\Delta : t \geq 0\}$ must exit and subsequently never return to C.

(Item 2 $\Rightarrow$ Item 3.) Instantiate Item 2 with $\ell = \ell_{\log}$ and $\varepsilon = 1$, and let the corresponding b and N be given. Pick any minimizing sequence $\{\Delta_i^{(1)}\}_{i=1}^{\infty}$ for $L(\cdot; D)$, meaning $L(A_i^{(1)}; D) \to L(A; D)$. Next produce a subsequence $\{\Delta_i^{(2)}\}_{i=1}^{\infty}$ of $\{\Delta_i^{(1)}\}_{i=1}^{\infty}$ by removing all elements with $L(A_i^{(1)}; D) > L(A; D) + 1$ (this procedure must be possible, since otherwise $\{\Delta_i^{(1)}\}_{i=1}^{\infty}$ is not a minimizing sequence). By the assumed properties of Item 2, a sequence $\{\Delta_i^{(3)}\}_{i=1}^{\infty}$ may be produced from $\{\Delta_i^{(2)}\}_{i=1}^{\infty}$ where $k_i^{(3)} k_1 \leq b$, and $H_i^{(2)} = H_i^{(3)}$ over $D \cap N$, which in particular means $\{\Delta_i^{(3)}\}_{i=1}^{\infty}$ is also a minimizing sequence for $L(\cdot; D)$. But this is now a minimizing sequence lying within the l_1 ball in R^n of radius b, a compact set, thus some subsequence $\{\Delta_i^{(4)}\}$ has a limit $\Delta \in R^n$. Since $\exists \int \ell(y(H\cdot)x)\, d\mu_D$ is lower semi-continuous by Lemma 4.1.2 and finite everywhere on R^n (and thus continuous), it follows that Δ attains the desired infimal value, meaning $L(\Delta; D) = L(A; D)$.

Applying the duality relation in Chapter 4 to $L(\cdot; D)$ (i.e., applying Lemma 4.1.1 to $\int \ell \, d\mu_D$), the existence of a primal minimum Δ grants the existence of a dual maximum p satisfying $p \in D(H; \mu_D)$, and moreover

$$p(x;y) \in \partial(\ell(y(H\Delta)x)) = \{\nabla \ell_{\log}(y(H\Delta)x)\} = \frac{1}{1 + \exp(y(H\Delta)x)} \quad \text{for } \mu_D\text{-a.e. } (x;y).$$

As such, the choice $p^0(x;y) := 1/(1 + \exp(y(H\Delta)(x)))$ satisfies $p^0 := p \ \mu_D$-a.e., and thus $p^0 \in D(H; \mu_D)$; moreover $p^0 > 0$ everywhere, since $1/(1 + \exp(z)) > 0$ for all $z \in R$.

This reweighting p^0 was with respect to μ_D, so to finish, define $p(x;y) := p^0(x;y) 1[(x;y) \in D]$. By construction, $[p > 0] = D$. Finally, given any $\Delta \in R^n$,

$$\begin{aligned} \int y(H\Delta)(x) p(x;y)\, d\mu(x;y) &= \int y(H\Delta)(x) p^0(x;y) 1[(x;y) \in D]\, d\mu(x;y) \\ &= \int y(H\Delta)(x) p^0(x;y)\, d\mu_D(x;y) \\ &= 0 : \end{aligned}$$

It follows that $p \in D(H, \ell)$, and that $D \in S_D(H, \ell)$.

(Item 3 $\Rightarrow$ Item 1.) Let $p \in D(H, \ell)$ with $D = [p > 0]$ be given, and take any $\in R^n$ satisfying $(D \setminus [y(H \ )x > 0]) > 0$. But notice then, since p decorrelates H ,

$$\begin{aligned} 0 &= \int p(x,y)y(H \ )(x)d \ (x,y) \\ &= \int_{D;y(H \ )(x)>0} p(x,y)y(H \ )(x)d \ (x,y) + \int_{D;y(H \ )(x)<0} p(x,y)y(H \ )(x)d \ (x,y): \end{aligned}$$

From this it follows that

$$\int_{D;y(H \ )(x)<0} p(x,y)y(H \ )(x)d \ (x,y) = \int_{D;y(H \ )(x)>0} p(x,y)y(H \ )(x)d \ (x,y) > 0;$$

where the inequality follows from $(D \setminus [y(H \ )(x) > 0]) > 0$ (Folland, 1999, Proposition 2.23(b)). The result follows. □

Appendix 5.C Deferred Material from Section 5.2

Proof of Proposition 5.2.1. As stated in the proposition, set $X = [\ 1; +1]^2$, and H to be the two projection maps $h_1(x) = x_1$ and $h_2(x) = x_2$. Next de ne a set of positive instances $\{p_i\}_{i=1}^{1}$, and their corresponding probability mass:

$$p_i = \begin{matrix} 1 \\ 0{:}5 \ 4^{2 \ i} \\ 1 \end{matrix} ; \qquad (p_i) = 2^{\ i \ 1}:$$

Here are the negative instances:

$$n_i = \begin{matrix} & 1 \\ 1 & 0{:}3 \ 4^{2 \ i} \end{matrix} ; \qquad (n_i) = 2^{\ i \ 1}:$$

Notice that has countable support, and $(X) = 1$. Furthermore, the vector $= (\ 1; +1)$ is a perfect separator: given any positive example p_i, $(H \ )(p_i) > 0$, and given negative example n_i, $(H \ )(n_i) < 0$. Note however that, as required by the proposition statement, the margins go to zero. However, given any $` \in L_c$, since $\lim_{z \to 1} `(z) = 0$,

$$0 \quad \inf L(A \ ) \quad \lim_{c \to 1} \int `(\ y_i(H(c \ ))_{z_i})d \ (z_i, y_i) = 0:$$

The key property of this construction is that the positive and negative examples are staggered; this will cause max margin solutions to avoid . As such, let any nite sample of size m be given. If all drawn examples have the same class y, then $\hat{} = (1 \quad y; 1 + y)$ (which is a maximum margin solution) has either n_1 or p_1 on the wrong side of the separator, and by choosing $c > 0$ large enough, $L(cA\hat{}) > b$ and $\hat{L}^b(cA\hat{}) < 1=b$.

As such, henceforth suppose there is at least one positive example, and at least one negative example. Suppose j and k respectively denote a sampled positive point p_j and sampled negative point n_k having highest index among positive and negative examples; these maxima exist since m is nite.

Every max margin solution is determine solely by p_j and n_k. To obtain one of them, de ne

$$:= \begin{bmatrix} (1+(n_k)_2)=(2+(p_j)_1+(n_k)_2) \\ (1+(p_j)_1)=(2+(p_j)_1+(n_k)_2) \end{bmatrix}:$$

To verify that this is a max margin solution, note that for any sampled (positive or negative) point z_i with label $y_i \; 2 \; f \quad 1; +1 g$,

$$y_i(H \quad)z_i \quad (H \quad)(p_j) = \quad (H \quad)(n_k) = \quad h \; ; n_k i = \frac{(p_j)_1(n_k)_2}{2 + (p_j)_1 + (n_k)_2} > 0:$$

By construction, however, $(p_j)_1 \; 6= \; (n_k)_2$, meaning is not a rescaling of . As such, is wrong for either all large p_i or n_i, and taking $\hat{} = q$ with q large, it follows that $L(A\hat{}) > b$ and $\hat{L}^b(A\hat{}) < 1=b$. □

Appendix 5.D Deferred Material from Section 5.3

Throughout this section, the following notation for measures will be employed

De nition 5.D.1. Given a nite measure and a set P, let $_P$ be the restriction of to P: for any measurable set S, $_P(S) = (P \setminus S)$. Note also that $d_P(x; y) = 1((x; y) \; 2 \; P)d(x; y)$.

5.D.1 Primal Hard Cores

In light of the duality relationship for L (cf. Lemma 4.1.1), the de nition for hard cores, provided in Section 5.3, is tied to the convex dual to L. Analogously, it is possibly to de ne a primal form of hard cores, which will be lead to a proof of Theorem 5.3.1.

De nition 5.D.2. De ne $S_P(H;)$ to contain all sets C for which there exists a sequence $\{ _i\}_{i=1}^{1}$ satisfying the following properties.

1. Every $_i$ and $(x; y) \in C$ satis es $y(H _i)x = 0$.
2. For -almost-every $(x; y)$ in C^c, $y(H _i)x \" 1$.

A primal hard core P is a (-a.e.) minimal set within $S_P(H;)$:

$$P \in S_P(H;) \quad \text{and} \quad \forall C \in S_P(H;) \quad (P \cap C) = 0 :$$

Lemma 5.D.3. $S_P(H;)$ is closed under countable intersections.

Proof. To start, note that $S_P(H;)$ is closed under nite intersections as follows. Let $\{C_i\}_{i=1}^{p}$ be given with $C_i \in S_P(H;)$, and let corresponding sequences $\{ _j^{(i)}\}_{j=1}^{1}$ be given as granted by the de nition of $S_P(H;)$. De ne $C := \cap C_i$ and $_j := \sum_i {}_j^{(i)}$. By construction, for every $(x; y) \in C$ and pair $(i; j)$, $y(H _j^{(i)})x = 0$, and thus $y(H _j)x = 0$. Next, for each C_i, de ne $C_i^0 \subseteq C_i^c$ with $(C_i^0) = (C_i^c)$ so that, for every $(x; y) \in C_i^0$, $y(H _j^{(i)})x \" 1$. Correspondingly, de ne $C^0 := \cup_i C_i^0$, where $(C^0) = (C^c)$. Now let any $(x; y) \in C^0$ and any $B > 0$ be given. For each i, there are two cases: either this is an area where $y(H _j^{(i)})x \" 1$, or $y(H _j^{(i)})x = 0$. In the rst case, let T_i denote an integer, as granted by $y(H _j^{(i)})x \" 1$, so that for all $j \quad T_i$, $y(H _j^{(i)})x > B$. For those i where $(x; y) \notin C_i^0$ (but still $(x; y) \in C^0$), due to the ruled out nullsets, $y(H _j^{(i)})x = 0$, safely set $T_i = 0$. To nish, taking $T := \max_i T_i$, it follows that for every $j > T$, $y(H _j)x > B$, whereby it follows that $y(H _j)x \" 1$ over C^0, and thus over C^c -a.e.

Now let a countable family $\{D_i\}_{i=1}^{1}$ be given with $D_i \in S_P(H;)$, and de ne $D = \cap_i D_i$. Consider the optimization problem

$$p := \inf \int \exp(\ y(H)x) d _{D^c}(x; y) : \ \in R^n; \forall (x; y) \in D \ y(H)x = 0 \ :$$

De ne $E_j := \cap_{i \ j} D_i$, whereby $D := \cap_j E_j$. Since $(X \ f \ 1; +1g) < 1$, by continuity of measures from above (Folland, 1999, Theorem 1.8(d)), for any > 0 there exists E_k with $(D) > (E_k)$. Since it was shown above that $S_P(H;)$ is closed under nite intersections, $E_k = \cap_{i \ k} D_i \in S_P(H;)$; consequently, let $\{ _i\}_{i=1}^{1}$ to be a sequence of predictors certifying that $E_k \in S_P(H;)$, as according to the de nition; since $E_k \quad D$, it follows that $y(H _i)x = 0$ over

E_k and thus over D as well. It follows by the dominated convergence theorem (with dominating function $(x; y) \mapsto \exp(0)$) that

$$p \lim_{i \to 1} \int \exp(-\gamma(H_i)x)\, d_{D^c}(x; y) \leq 0 + \exp(0)\, (E_k \cap D) = \mu(E_k) \leq \mu(D) < \infty.$$

Since μ was arbitrary, it follows that $p = 0$.

As such, for any $n \in Z_{++}$, choose $\gamma_n \in R^n$ with $\gamma(H_n)x = 0$ over D satisfying

$$\int \exp(-\gamma(H_n)x)\, d_{D^c}(x; y) < 1/n^2.$$

By Markov's inequality, it follows that

$$_{D^c}([\exp(-\gamma(H_n)x) \geq 1/n]) \leq n \int \exp(-\gamma(H_n)x)\, d_{D^c}(x; y) < 1/n.$$

Since $\exp(-\gamma(H_n)x) = \exp(0) = 1$ over D, it follows that $f(x; y) \mapsto \exp(-\gamma(H_n)(x))\}_{n=1}^{1}$ is Cauchy in measure, and moreover converges in measure to the function $(x; y) \mapsto 1[(x; y) \in D]$. Consequently, there exists a subsequence $\{\gamma_i\}$ with $\exp(-\gamma(H_i)x) \to 1[(x; y) \in D]$ μ-a.e. (Folland, 1999, Theorem 2.30). This is only possible if $\gamma(H_i)x \to \infty$ for μ-a.e. $(x; y) \in D^c$, and the result follows, with $\{\gamma_i\}_{i=1}^{1}$ as the certifying sequence for D, since every $\gamma(H_i)x = 0$ for $(x; y) \in D$ by construction. □

Theorem 5.D.4. *Every finite hypothesis class H and probability measure μ admits a primal hard core.*

Proof. Consider the optimization problem

$$p := \inf\{\mu(C) : C \in S_P(H; \mu)\}.$$

Since S_P is nonempty (it always contains $X \cup \{-1; +1\}$ with certifying sequence $\gamma_i = 0$ for every i) and μ is a finite nonnegative measure, the infimum is finite. Let $\{C_i\}_{i=1}^{1}$ be a minimizing sequence, and define $D_j := \cap_{i \leq j} C_i$ and $D := \cap_{j=1}^{1} D_j = \cap_{i=1}^{1} C_i$. By Lemma 5.D.3, $D_j \in S_P(H; \mu)$ for every j, and since $\mu(D_j) \leq \mu(C_j)$, it follows that $\{D_j\}_{j=1}^{1}$ must also be a minimizing sequence to the above infimum. Finally, since μ is finite and Lemma 5.D.3 also grants $D \in S_P(H; \mu)$, then

by continuity of measures from above (Folland, 1999, Theorem 1.8(d)),

$$(D) = \lim_{j \to \infty} (D_j) = p.$$

Since $D \in S_P(H, \ )$ attains the infimum, it is a primal hard core. □

With existence of primal hard cores out of the way, the next key is the equivalence to (dual) hard cores.

Theorem 5.D.5. Let finite hypothesis class H, probability measure $\ $, a (dual) hard core C, as well as a primal hard core P be given. Then C and P agree on all but a null set.

The proof uses the following technical lemma.

Lemma 5.D.6. Let finite hypothesis class H, probability measure $\ $, $C_1 \in S_P(H, \ )$, as well as a $\ _2 \in \mathbb{R}^n$ be given, with $y(H \ _2)x \geq 0$ for $(x, y) \in C_1$ (but potentially $y(H \ _2)x < 0$ elsewhere). Then $C_1 \cap [y(H \ _2)x > 0] \in S_P(H, \ )$.

Proof. Let $C_1, \ _2$ be given as specified. Let $\{ {}^{(1)}_i \}_{i=1}^{\infty}$ be a certifying sequence for C_1. Define $P := [y(H \ _2)x > 0]$ and $C_3 := C_1 \cap P = C_1 \cap [y(H \ _2)x > 0]$; the goal is to show $C_3 \in S_P(H, \ )$.

Let $i \in Z_{++}$ be arbitrary; the following steps will construct ${}^{(4)}_i$, a certifying sequence for C_3, meaning $C_3 \in S_P(H, \ )$.

First, let c be sufficiently large so that ${}^{(2)}_i := c \ _2$ satisfies

$$\int \exp(\ y(H \ {}^{(2)}_i)x) \ _P(x, y) < 1/i^2.$$

By Markov's inequality, it follows that

$$_P([\exp(\ y(H \ _2)x) \geq 1/i]) \leq i \int \exp(\ y(H \ _2)x) \ _P(x, y) < 1/i. \qquad (5.D.7)$$

Consequently define $P_i := [y(H \ _2)x > \ln(i)]$, where the above statements show $(P_i) > (P) - 1/i$.

Next, since $\exp(\ y(H \ {}^{(1)}_i)x) \to 1[(x, y) \in C_1]$ -a.e. and $(X \times \{-1, +1\}) < 1$, by Egoroff's Theorem (Folland, 1999, Theorem 2.33), this convergence is uniform over a subset $S_i \subseteq (X \times \{-1, +1\})$ with $(S_i) > (X \times \{-1, +1\}) - 1/i$. In particular, there exists an integer

T_i so that, for any $(x;y) \in S_i \setminus C_1$,

$$y(H^{(1)}_{T_i})x > k^{(2)}_i k_1 + \ln(i):$$

As such, define $^{(3)}_i := {}^{(1)}_{T_i} + {}^{(2)}_i$. First, for any $(x;y) \in C_3$ and any i,

$$y(H^{(3)}_i)x = 0 = y(H^{(1)}_i)x = y(H_2)x:$$

On the other hand, for any $(x;y) \in S_i \setminus C_1$, since $\sup_{n;x} |y_n(x)| \leq 1$,

$$\begin{aligned} y(H^{(3)}_i)x &= y(H^{(1)}_{T_i})x + y(H^{(2)}_i)x \\ &> k^{(2)}_i k_1 + \ln(i) - k^{(2)}_i k_1 = \ln(i): \end{aligned}$$

Lastly, as shown above, for any $(x;y) \in P_i$,

$$y(H^{(3)}_i)x = 0 + y(H^{(2)}_i)x \geq \ln(i):$$

Combining the above facts,

$$([|\exp(-y(H^{(3)}_i)x) - 1[(x;y) \in C_3]| \geq 1=i]) < \mu(C_1^c \cap S_i) + \mu(P \cap P_i) \leq 2=i:$$

It follows that $\exp(-y(H^{(3)}_i)x) \to 1[(x;y) \in C_3]$ in measure (and the sequence of functions is Cauchy in measure), and thus there is a subsequence $\{^{(4)}_i\}_{i=1}^{1}$ for which $\exp(-y(H^{(4)}_i)x)$ converges to $1[(x;y) \in C_3]$ μ-a.e. (Folland, 1999, Theorem 2.30). It follows that $\{^{(4)}_i\}_{i=1}^{1}$ is the desired sequence certifying that $C_3 \in S_P(H;\mu)$. □

Proof of Theorem 5.D.5. If $\mu(P \cap C) > 0$, then by the maximality of C, P is a set of positive measure away from any element of $S_D(H;\mu)$, and in particular $P \notin S_D(H;\mu)$, and thus Theorem 5.B.1 provides the existence of $\in \mathbb{R}^n$ with $\mu(P \setminus [y(H)x \leq 0]) = \mu(P)$ and $\mu(P \setminus [y(H)x > 0]) > 0$. But then, by Lemma 5.D.6, $P' := P \cap [y(J)x > 0]$ has $\mu(P') < \mu(P)$ but $P' \in S_P(H;\mu)$, which contradicts the μ-a.e. minimality of hard core P.

Now suppose $\mu(C \cap P) > 0$, and let μ_C denote the restriction of μ to C: for any C,

$_C(C) := (C \setminus C)$. Consider the optimization problem

$$\inf \int \exp(-y(H\,)(x))\, d\,_C(x,y) : \in R^n \;:$$

Consider the sublevel set of 1-suboptimal points for this problem. By Theorem 5.B.1, there exists B so that each $$ in this sublevel set has ^{0} with $H\, = H\,^0$ -a.e. and $k\,^0k_1 \le B$. However, by the definition of P, there exists a sequence $\{\,_i\}_{i=1}^1$ which is zero over P and approaches 1 -a.e. over P^c, and in particular over the positive measure set $C \cap P$. Thus, taking any $$ in the 1-suboptimal set, notice that

$$\lim_{i\to 1} \int \exp(-y(H(\, + \,_i))x))\, d\,_C(x,y) = \int \exp(-y(H\,)(x))\, 1[(x,y) \notin P]\, d\,_C(x,y) =: p.$$

Since $$ has a bounded representation, $\exp(-y(H\,)x) \ge 0$, which combined with $(C \cap P) > 0$ grants $p < \int \exp(-y(H\,)(x))\, d\,_C(x,y)$ (Folland, 1999, Theorem 2.23(b)). But since $\to \int \exp(-y(H\,)(x))\, d\,(x,y)$ is continuous in $$ for any finite measure $$ (cf. Lemma 5.A.2), there must exist a large j so that

$$\int \exp(-y(H\, + \,_j)(x))\, d\,_C(x,y) < \int \exp(-y(H\,)(x))\, d\,_C(x,y);$$

and moreover $y(H(\, + \,_j))(x) > B$ for a subset of C with positive measure. But that means $\, + \,_j$ is in the 1-sublevel set (since it beats $$, which was chosen to reside in the 1-sublevel set), but can not have a representation with norm at most B (since H is a bounded linear operator with operator norm 1), contradicting Theorem 5.B.1. □

5.D.2 Proof of Theorem 5.3.1

This is now just a consequence of the equivalence to primal hard cores, and the structure over C developed in Theorem 5.B.1 (which was used to prove the equivalence to primal hard cores as well).

Proof of Theorem 5.3.1. The second property is direct from Theorem 5.B.1. For the first property, since primal hard cores exist and are -a.e. equivalent to (dual) hard cores (cf. Theorem 5.D.5), the statement thus follows by taking the sequence provided by the definition of any primal hard core. □

Appendix 5.E Deferred Material from Section 5.4

Proof of Theorem 5.4.1. (Item 1) Let $\{H_i\}_{i=1}^{\infty}$ be given as per Theorem 5.3.1. Automatically, $y(H_i)x = 0$ for $(x,y) \in C$. And since $y'(H_i)x' \leq 1$ for μ-a.e. $(x',y') \in C^c$, it follows from the definition of L_c that $\lim_{i\to\infty} \ell(y'(H_i)x) = 0$.

(Item 2) This is a consequence of Theorem 5.B.1. □

Proof of Theorem 5.4.2. (Item 1) Let a sequence $\{H_i\}_{i=1}^{\infty}$ be given as provided by Theorem 5.3.1. In particular, $\exp(-y(H_i)x) \to 1[(x,y) \in C]$ μ-a.e.. Now choose a finite sample size m; by Egoroff's Theorem (Folland, 1999, Theorem 2.33), for any $\gamma > 0$, there exists S with $\mu(S) > \mu(X \times \{-1,+1\}) - \gamma/m$ over which this convergence is uniform. As such, choose α so that $\exp(-y(H_\alpha)x) < 1/2$ over $S \cap C^c$, meaning in particular $y(H_\alpha)x > 0$ for every $(x,y) \in S \cap C^c$. The probability over a draw of m points that some within C^c are misclassified by H_α has upper bound bound

$$\Pr[\exists (x_i,y_i) \in C^c \;\; y(H_\alpha)x \leq 0] \leq m\,(\mu(C^c) - \mu([y(H_\alpha)x \leq 0])) < \gamma.$$

Since γ can be made arbitrarily small, the probability of failure is zero. Furthermore, since H_α satisfies $y(H_\alpha)(x) = 0$ μ-a.e. over C (cf. Theorem 5.3.1), it also follows that, with probability 1, H_α abstains on every example falling within C.

(Item 2) Let $\gamma > 0$ and $\ell \in L_c$ be given. Choose $b > 0$, as provided by Theorem 5.4.1, so that every $H \in \mathbb{R}^n$ with $L(A_\ell;C) \leq L(A_\ell;C) + 4 + \gamma$ has a representation H' with $\|H'\|_1 \leq b$, where $H = H'$ everywhere along $C \setminus N$, where $\mu(N) = 0$; henceforth, discard the (measure zero) event that any example falls within N. Additionally, choose $c > 0$ as provided by Proposition 5.A.3 so that, given m_C i.i.d. points within $C \setminus N$, every $H \in \mathbb{R}^n$ with $\|H\|_1 \leq b$ satisfies, with probability at least $1 - \delta$,

$$|L(A_\ell;C) - \hat{L}(A_\ell;C)| \leq c'(C)\frac{\sqrt{\ln(n)} + \sqrt{\ln(2/\delta)}}{\sqrt{m_C}}; \tag{5.E.1}$$

where the extra $\mu(C)$ comes from applying Proposition 5.A.3 to μ conditioned on points falling within C, which can be related to the above left hand side via this normalization by $\mu(C)$. Henceforth, for simplicity, set $c := c'(C)$.

Now consider any $H \in \mathbb{R}^n$ with no representation $\|H'\|_1 \leq b$ so that $H = H'$ over $C \setminus N$, which directly entails, by Theorem 5.4.1, that $L(A_\ell;C) - L(A_\ell;C) > \gamma + 4$. Additionally choose

and any $2 R^n$ with $L(A ;C)$ $L (A ; C) < 1$, whereby the choice of $b > 0$ indicates that, without loss of generality, $k k_1$ b. Since $7! \int_{`}(A)d_C$ is continuous (cf. Lemma 5.A.2), considering the line segment $f +(1) : 2 [0;1]g$, there must exist $\hat{}$ along this line segment with with

$$+3 \quad L (A\hat{};C) \quad L (A ; C) \quad +4;$$

let $\hat{}^0$ be a representation with $k\hat{}^0k_1$ b and $H\hat{} = H\hat{}^0$ over $C \setminus N$ (and thus it holds for every example). Applying the deviation inequality in eq. (5.E.1) twice,

$$\begin{aligned}
\hat{L}_b(A\hat{};C) \quad \hat{L}_b(A ;C) \quad L (A\hat{}^0;C) \quad L (A ;C) \quad & 2c\frac{\sqrt{\ln(n)}+\sqrt{\ln(2=)}}{\sqrt{m_C}}: \\
= L(A\hat{}^0;C) \quad L (A ; C) \quad & (L(A ;C) \quad L (A ; C)) \\
& 2c\frac{\sqrt{\ln(n)}+\sqrt{\ln(2=)}}{\sqrt{m_C}}: \\
> (\quad +3) \quad (1) \quad & 2c\frac{\sqrt{\ln(n)}+\sqrt{\ln(2=)}}{\sqrt{m_C}}: \\
& ;
\end{aligned}$$

where the last step used the lower bound on m_C. Returning to $2 R^n$ as speci ed above, convexity, in the form of Lemma 5.A.5, grants that $\hat{L}_b(A ;C) < \hat{L}_b(A\hat{};C)$ implies $\hat{L}_b(A\hat{};C)$ $\hat{L}_b(A ;C)$, and thus

$$\hat{L}_b(A ;C) \quad \hat{L}_b(A ; C) \quad \hat{L}_b(A\hat{};C) \quad \hat{L}_b(A ;C) > :$$

Since was arbitrary, it follows that every with no representation $k \ ^0k_1 > b$ that has agreement of H and $H^{\ 0}$ -a.e. over C does not lie in the empirical -sublevel set. Since $\hat{L}_b(;C)$ is convex and continuous (cf. Lemma 5.A.2), the -sublevel set is nonempty, and thus every ^{0} within it has a representation $k \ ^{00}k_1$ b. □

Appendix 5.F Deferred Material from Section 5.5

Proof of Proposition 5.5.2. This proof is essentially a repackaging of various results and comments due to Bartlett et al. (2006). Fix any $` 2 L_c$; $`$ is convex, increasing at 0, and di erentiable at 0, which grants that the corresponding -transform is classi cation calibrated (Bartlett et al., 2006, Theorem 6, although note losses in this thesis are increasing rather than decreasing). It follows

that $(R(f) - R(F)) \le L(f) - L(F_Y)$; (Bartlett et al., 2006, Theorem 3, part 3(c)).

Next, $(0) = 0$ (Bartlett et al., 2006, Lemma 5, part 8), $(r) > 0$ when $r > 0$ (Bartlett et al., 2006, Lemma 5, part 9(b)), and since $$ is convex by construction (Bartlett et al., 2006, De nition 2), it follows by Lemma 5.A.5 that $$ is increasing. Since $$ is continuous as well, (Bartlett et al., 2006, Lemma 5, part 6), it follows that $$ has a well-de ned inverse along the image $([0;1])$. Finally, the fact that $^{-1}(r) \to 0$ as $r \to 0$ is due to Bartlett et al. (2006, Theorem 3, part 3(b)). □

Proof of Theorem 5.5.4. Throughout this proof, $^0 := /8$ will be the failure probability of various crucial events; the nal statement is obtained by unioning them together, and subsequently throwing them all out. Note also that some of the statements vacuously hold if $(C) = 0$ or $(C) = (X \setminus \{-1,+1\})$ (i.e., when terms depending on either appear in denominators); interpret these expressions as simply being 1, whereby the bounds hold automatically.

(Item 1) Let S_C and S_+ respectively denote the set of samples landing in C and C^c, where the notation proposed in the theorem statement provides $m_C = |S_C|$ and $m_+ = |S_+|$. By a Cherno bound (Kearns and Vazirani, 1994, Theorem 9.2), basic deviations for these quantities are

$$\Pr[|S_C| < (\ (C)\ )m] \le \exp(-m\ ^2/2) = {}^0/4,$$
$$\Pr[|S_+| < (\ (C^c)\ )m] \le \exp(-m\ ^2/2) = {}^0/4,$$

where $= \sqrt{\frac{1}{2m}\ln\frac{1}{^0}}$, and Pr denotes the product measure corresponding to m copies of . Label these failure events F_1 and F_2, and henceforth rule them out.

(Item 2) As provided by Theorem 5.4.2, there exists $\in R^n$ with $y_i(Hx_i) > 0$ for all (x_i, y_i) falling in C^c, and $y_i(Hx_i) = 0$ for those landing in C. Consequently,

$$\hat{L}(A) = \inf_{c>0} \hat{L}(A(\ + c\ ); C) + \hat{L}(A(\ + c\ ); C^c)$$
$$= \inf_{c>0} \hat{L}(A\ ; C)$$
$$\hat{L}(A)\ :$$

Combining this with

$$\hat{L}(A\ ; C^c) + \hat{L}(A\ ; C) = \hat{L}(A\ ) \quad \hat{L}(A\ ) + \quad ;$$

it follows that

$$\hat{L}(A\ ; C^c) \quad \hat{L}(A\ ) \quad \hat{L}(A\ ; C) + \ = \hat{L}(A\ ;\ C) \quad \hat{L}(A\ ; C) + \quad :$$

Next, since $\grave{}(0) > 0$ and $\grave{}$ is nondecreasing (cf. Lemma 5.A.1),

$$\hat{R}(H\ ; C^c) \quad \frac{\hat{L}(A\ ; C^c)}{\grave{}(0)} = \frac{}{\grave{}(0)} :$$

To obtain eq. (5.5.5) from here, first notice that S_+, the portion of the sample falling within C^c, can be interpreted as an i.i.d. sample from the probability measure $(\ \setminus\ C)=\ (C)$. Next, the VC dimension of span(H) is the VC dimension of linear separators over the transformed space

$$f((h_1(x); h_2(x); : : : ; h_n(x)); y) : (x; y) \ 2\ X\ \ f\ \ \ 1; +1 gg;$$

namely, it is n. As such, eq. (5.5.5) follows by an application of a relative deviation version of the VC Theorem (Boucheron et al., 2005a, discussion preceding Corollary 5.2). albeit with an adjustment to apply this uniform deviation bound with conditioned on C^c, and then rescaling to remove this conditioning (just as in the proof of Theorem 5.4.2).

To obtain eq. (5.5.6), note that $< \grave{}\ (0)/m$ means there are no mistakes over C^c:

$$\begin{aligned} \grave{}(0) > m \quad & m_+ \hat{L}(A\ ; C^c) \\ & \sum_{i=1}^{m_+} \grave{}(\ y_i(H\ )x_i) \\ & \max_{i\ 2\ [m_+]} \grave{}(\ y_i(H\ )x_i); \end{aligned}$$

that is to say, since $\grave{}$ is nondecreasing, for every $(x_i ; y_i)\ 2\ S_+$, $0 < y_i(H\ )x_i$. Plugging $\hat{R}(H\ ) = 0$ into the same relative deviation bound as before (Boucheron et al., 2005a, discussion preceding Corollary 5.2), the second bound follows.

(Item 3) By Theorem 5.4.2, there exist constants $b > 0$ and c (b), depending on

$H; ; ; C$, so that with probability at least $1 \quad {}^{0}$, if $m_C \quad c^2(\ln(n) + \ln(1=\ {}^{0}))$, then every $$-suboptimal predictor over C, and in particular $$, has a representation ^{0} which is equivalent to $$-a.e. over C, and satisfies $k\ {}^{0}k_1 \quad b$. As such, since

$$\underline{b}(A\ ; C) = \underline{b}(A\ {}^{0}; C) \qquad \text{and} \qquad \underline{b}(A\ ; C) = \underline{b}(A\ {}^{0}; C);$$

an application of Proposition 5.A.3 grants

$$\begin{aligned} L(A\ ; C) &= L(A\ {}^{0}; C) \\ &\quad \underline{b}(A\ {}^{0}; C) + \frac{c\ \sqrt{\ln(n)} + \sqrt{\ln(2=\ {}^{0})}}{\sqrt{m_C}} \\ &= \underline{b}(A\ ; C) + \frac{c\ \sqrt{\ln(n)} + \sqrt{\ln(2=\ {}^{0})}}{\sqrt{m_C}} \\ &\quad \underline{b}(A\ ; C) + \quad + \frac{c\ \sqrt{\ln(n)} + \sqrt{\ln(2=\ {}^{0})}}{\sqrt{m_C}}; \end{aligned}$$

where as in the proof of Theorem 5.4.2, the constant c also captures a correction since Proposition 5.A.3 is applied to $$ condition on points falling in C. Next, noting that Theorem 5.4.1 provides that a minimizing sequence to $L(A\ ; C)$ can be taken without loss of generality to lie within a compact set (e.g., points with l_1 norm at most b); combined with the continuity of $7! \int `(A\ )d\ _C$ (cf. Lemma 5.A.2), it follows that a minimizer $$ exists. By an application of McDiarmid's inequality, with probability at least $1 \quad {}^{0}$,

$$\underline{b}(A\ ; C) \quad \underline{b}(A\ ; C) \quad L\ (A\ ; C) + c\ \sqrt{\frac{2\ln(1=\ {}^{0})}{m_C}}:$$

(Note, $$ is independent of the sample, thus McDiarmid suces, with constant c (b) since $$ is in this initial sublevel set.) Combining these two pieces, it follows that

$$L(A\ ; C) \quad L\ (A\ ; C) \qquad + \frac{c\ \sqrt{\ln(n)} + 4\sqrt{\ln(2=\ {}^{0})}}{\sqrt{m_C}};$$

which is precisely eq. (5.5.7).

To produce eq. (5.5.8), the definition of the $$-transform (cf. Proposition 5.5.2), combined

with Equation (5.5.7), provides

$$\begin{aligned} R(H\ ;\ C)\ \ R\ (F;C) &\quad {}^{1}(L(A\ ;C)\ \ L\ (F_Y;C)) \\ = &\quad {}^{1}(L(A\ ;C)\ \ L\ (A\ ;\ C) + L(A\ ;\ C)\ \ L\ (F_Y;C)) \\ &\quad {}^{1}@\ +\ \frac{c\ \sqrt{\ln(n)} + 4\ \sqrt{\ln(2=\ ^{0})}}{\sqrt{m_C}} + L(A\ ;\ C)\ \ L\ (F_Y;C)A\ : \end{aligned}$$

(Item 4) Combining the lower bound on m with Item 1,

$$\begin{aligned} m_+ &\quad m\ (C^c)=2; \\ m_C &\quad m\ (C)=2\quad c^2(\ln(n) + \ln(1 = \ ^{0})); \end{aligned}$$

the first two bounds will allow expressions to be simplified, whereas the last bound will allow an invocation of item 3.

As such, combining all preceding bounds (and making use of the refinement over C^c when $< \ ` (0)=m$),

$$\begin{aligned} R(H\ )\ \ R\ (F) = (\ R(H\ ;C)\ \ R\ (F;C)) + (\ R(H\ ;C^c)\ \ R\ (F;C^c)) \\ \underbrace{\qquad}_{=0} \\ {}^{1}@\ +\ \frac{c\ \sqrt{\ln(n)} + 4\ \sqrt{\ln(2=\ ^{0})}}{\sqrt{m_C}} + L(A\ ;\ C)\ \ L\ (F_Y;C)A \\ +\ \ (C^c)\frac{4(n\ln(2m_+\ + 1) + \ln(4\ = \ ^{0})}{m_+} \\ {}^{1}@\ +\ \ (C^c)\frac{c\ \sqrt{2}\ \sqrt{\ln(n)} + 4\ \sqrt{\ln(2=\ ^{0})}}{\sqrt{m\ (C)}} + L(A\ ;\ C)\ \ L\ (F_Y;C)A \\ +\ \frac{8(n\ln(m\ \ (C^c) + 1) + \ln(4\ = \ ^{0})}{m}: \end{aligned}$$

□

Appendix 5.G Bibliographic Notes

As discussed at the beginning of the chapter, there are a variety of ways to produce general statistical guarantees for boosting methods; results of this type will be discussed in more

detail in the bibliographic remarks in Chapter 6.

Focusing then on the nite-dimensional case with a goal of handling general losses, there are very few results, since the problem is not regularized or constrained in any way. Similarly, there are statistical guarantees in the presences of margins bounded below by a positive constant (Schapire et al., 1997), but this of course does not hold in general. If the goal is relaxed to consistency and just handling the logistic loss, the only results appear to be those related to the consistency of maximum likelihood (Lebanon, 2008, Gourieroux and Monfort, 1981), which require a number of conditions that may fail in general (in particular, the in mum of the optimization problem may fail to exist, which corresponds to a correct model failing to exist).

Note that discussion on hard cores and some of their history may be found in Section 1.E.

Acknowledgements

This chapter is based on work by the dissertation author which is currently under preparation.

Chapter 6

In nite Hypothesis Classes

This chapter provides numerical and statistical guarantees while allowing both the support of and the size ofH to be in nite; however, the guarantees only hold for losses which, like the logistic loss, are Lipschitz.

As has been the case throughout this thesis, the strategy is to investigate the structure imposed on the optimization problem by and H, and thereafter provide the numerical and statistical guarantees.

Unlike the preceding chapters, hard cores are not used, but the structure is similar. The rst case is when the optimal convex risk over the distribution, L(A), is zero, which is analogous to the case that all hard cores have measure zero (cf. Proposition 6.3.6). In this situation, the analysis exhibits a quantity similar to the weak learning rate which can once again establish quick decrease of the primal objective function. On the other hand, whenL(A) > 0, the analysis shows that every sample with high probability has some tricky examples, which in turn introduce curvature and constrain the norms of the chosen predictors; from here, a convergence analysis follows readily.

6.1 Overview

This chapter will establish statistical properties of Boost for any of the three line searches, together with the second stopping condition which ensures roughly m^a iterations occur. The basic consistency guarantee, as well as an overview of the analysis, appear in Section 6.2. The class of loss functions to which the results apply is as follows.

De nition 6.1.1. Let L_{2d} denote twice continuously di erentiable convex losses. Additionally, let L_{lg} contain all di erentiable convex Lipschitz losses $\ell : \mathbb{R} \to \mathbb{R}_+$ with tightest Lipschitz

constant $_2 := \sup_{x \in y} j`(x) \quad `(y)j=jx \quad yj$ as follows. First, every $`$ has $`^0 2 [0; \ _2]$ everywhere and $`^0 2 [\ _1; \ _2]$ over $R_+$ for some $0< \ _1 \ _2$. Second, every $`$ has Lipschitz gradients with (tightest) parameter B_2, meaning $B_2 := \sup_{x;y \ 2 R}(`^0(y) \quad `^0(x))=(y \quad x) < 1$.

Although the most general guarantees require $` 2 L_{lg} \setminus L_{2d}$, the separable case needs only $` 2 L_{lg}$, which allows consideration of the interesting piecewise quadratic loss $`_{huss}(x)$, which was used by Impagliazzo (1995, Proof of Lemma 1) in the study of hard cores.

Section 6.3 considers the separable case $L(A) = 0$. Similarly to the analogous case in Chapter 5, the convex risk L is not controlled in this scenario, but rather only the classi cation risk R. The analysis has a minor handicap; rather than applying to the nal coe cients $_t$, it applies to the coe cient vector $^$, de ned as follows.

De nition 6.1.2. Given the sequence of coe cient vectors $f \ _i g_{i=1}^t$, let $^$ denote the vector achieving the best empirical classi cation risk, meaning $^ := \arg\min_{ } \hat{R}(H) : \ 2 f \ _i g_{i=1}^t g$.

While this additional indirection | which can be computed easily as the algorithm progresses | is perhaps unnecessary, it leads to a simple analysis in the case $L(A) = 0$. In light of the fact that similar structure arose in Chapter 5, it is interesting to wonder if, in this regime, boosting methods are in fact better adapted to R than to L.

In the separable case, the hypothesis class must have nite VC dimension, denoted as follows.

De nition 6.1.3. Let F_{vc} contain all classes H of nite VC dimension, denoted $V(H) < 1$.

The nonseparable follows in Section 6.4, and holds for losses within $L_{2d} \setminus L_{lg}$. Additionally, the hypothesis class must not only have nite VC dimension, indicating it is not too large, it must also satisfy the following condition which indicates it is not too small.

De nition 6.1.4. Let $F_{ds}( )$ contain every class H whose linear span $span(H)$ is dense (in the $L^1( )$ topology) in the collection of all bounded measurable functions over X.

Conditions similar to those de ning $F_{ds}( )$ are usually called dense class assumptions (Bartlett and Traskin, 2007, Condition 1, Denseness), or completeness assumptions (Breiman, 2000, De nition 1); for a more extensive discussion of these conditions, please see Section 6.A; for the time being, the important point is that reasonable elements of $F_{vc} \setminus F_{ds}( )$ exist; in particular, the following result provides that if $X = R^d$, then decision lists and decision trees with d axis-aligned splits su ce.

Proposition 6.1.5. Suppose $X = \mathbb{R}^d$, and let μ be a Borel probability measure over X. If $\mathrm{span}(H)$ contains all indicators of products of half-open intervals of the form $\prod_{i=1}^{d} [a_i, b_i)$, where $a_i < b_i$, then $H \in \mathcal{F}_{ds}(\mu)$.

As a final bit of notation, define the shorthands $L := L(A)$ and $R := R(H)$.

Lastly, note that this section will be forced to develop its own convergence analysis of Boost, and in particular can not use those present in Chapter 3. The reason is simply that the many "constants" in the bounds there are dependent on the sample, and more importantly do not concentrate with high probability in a way which allows those bounds to be used here.

6.2 Consistency Statement and Analysis Sketch

The analysis considers two cases: either $L = 0$ (separable) or $L > 0$ (nonseparable). (By Proposition 6.B.3, $L = 0$ implies finite samples have a separating choice $\in H$ almost surely.)

When the instance is separable, the improvement in objective value $L(A)$ in early iterations may be lower bounded by a margin-based quantity related to the classical weak learning rate; while this quantity is a random variable, with high probability it can be lower bounded by the analogous quantity over the distribution (which will be shown positive if the instance is separable). The bulk of the analysis is in constructing and controlling this quantity; the optimization and generalization analysis thereafter is straightforward, yielding a rate of roughly $O(1/m^{1/3})$ when $a = 2/3$.

When the instance is not separable, every weak learner makes a fair number of mistakes, and thus the algorithm makes more hesitant progress. Concretely, with high probability, the norms of the iterates are bounded, and moreover the quantity $m^{-1} \sum_i \ell''(A)_{x_i,y_i}$, which is roughly the Hessian in axis-aligned directions (and relevant to coordinate descent), is also lower bounded. This in turn allows adaptation of the optimization analysis due to Zhang and Yu (2005). While the rate in this case is still roughly $O(1/m^{1/c})$, unfortunately the exponent $c > 1$ depends both on ℓ and on H (but is of course finite).

As a final point of interest, each case, in order to respectively establish either fast decrease or the norm constraints, considers the behavior of the reweighted average margins

$$\int (A\theta)\, d\pi = \int (A\theta)_{x,y}\, p(x,y)\, d\mu(x,y); \tag{6.2.1}$$

where $\in$ and $p \in L^1(\ )$. In the separable case, this quantity is studied for a single good as p varies, whereas the nonseparable case studies a single bad p as varies.

Combining these nite sample results with the Borel-Cantelli lemma gives the following.

Theorem 6.2.2. Let loss $\ell \in L_{lg} \cap L_{2d}$, probability measure over $X \times \{-1,+1\}$, binary class $H \in \mathcal{F}_{vc} \cap \mathcal{F}_{ds}(X)$, and any stopping parameter $a \in (0,1)$ be given. Let $\hat{}_m$ denote the output of Figure 1.3 when run on m examples, and let $R_?$ to denote the Bayes error rate. Then $R(A\hat{}_m) \to R_?$ almost surely as $m \to \infty$.

6.3 The Separable Case ($L = 0$)

The rates in the case $L = 0$ will depend on the following quantity ( ), which directly embeds the reweighted margin expression in eq. (6.2.1). This quantity in turn depends on a set of reweightings $D(\ )$ which is a slightly constrained version of the set $D(H, \ )$ used to de ne hard cores in Section 1.7.

De nition 6.3.1. Let be any probability measure over $X \times \{-1,+1\}$ (relevant choices are and b), and let $\in [0,1]$ be given. De ne a permissible set of densities (with respect to)

$$D(\ ) := \{p \in L^1(\ ) : p \geq 0 \text{ -a.e.}, \|p\|_1 = 1, \|p\|_\infty \leq 1/\ \},$$

with the convention $1/0 = \infty$ in the case $= 0$. Additionally de ne

$$(\ ) := \inf_{p \in D(\ )} \sup \int (A\ )p\, d\ : \in\ , \|\ \|_1 \leq 1\ .$$

(When is a discrete measure, ( ) is almost equivalent to AvgMin_k as developed by Shalev-Shwartz and Singer (2008, Section 4.1).)

This quantity will play a role analogous to the weak learning rate in AdaBoost, which guarantees the algorithm makes speedy progress in certain separable cases. The correspondence between these two quantities will occupy much of this section; but rst, note the primary guarantee in the separable case.

Theorem 6.3.2. Let $\ell \in L_{lg}$ (with parameters $_1$, $_2$, B_2) and any $H \in \mathcal{F}_{vc}$ be given, and suppose $L = 0$. Let any error tolerance $\in (0,1]$ and any con dence parameter $\in (0,1]$ be given, and

for convenience set $^0 := {}_1 = {}_2$; by these choices, ${}_0( ) > 0$. Suppose Figure 1.3 is run with stopping parameter $a \in (0;1)$, and the sample size m satisfies

$$m \quad \max\left(\frac{2}{({}_0( ))^2} \ln \frac{4}{-} \; ; \; \frac{24 B_2 \grave{}(0)}{({}_1 {}_0( ))^2} \right)^{1=a} :$$

Then, with probability at least 1 , the algorithm's output $\hat{}$ satisfies

$$R(H\hat{}) \quad + 2 \sqrt{\frac{(V(H)+1)\ln(2 em) + \ln(8 =)}{m^{1 \ a}}} + 4 \frac{(V(H)+1)\ln(2 em) + \ln(8 =)}{m^{1 \ a}} :$$

To simplify this bound, first note that, for the logistic loss, $B_2 = 1=4$, $_1 = 1=2$, and $_2 = 1$ (cf. Lemma 6.B.1). Ignoring these terms, as well as $\grave{}(0) = 1$, (which can be set to 1/2), and ${}_0( )$; the choices $a = 1=2$ and $:= O(m^{\ a=2}) = O(m^{\ 1=4})$ grant that $O(1= ^2)$ iterations suffice to achieve classification risk $O(1=\sqrt{m})$, whereas the choices $a = 2=3$ and $:= O(m^{\ a=2}) = O(m^{\ 1=3})$ provide that $O(1= ^3)$ suffice to achieve error $O(m^{\ 1=3})$.

6.3.1 The Quantity ( )

To develop the meaning and necessity of ( ), first recall the classical notion of weak learnability, as defined in Section 1.7. Both the weak learning rate and its empirical analogue (b) are close in definition of ${}_0( )$ and ${}_0(b)$; indeed, the latter subscripted quantities replace the inequalities and quantifiers in the former quantities with min a and suprema, which leads to the following correspondence.

Proposition 6.3.3. Let class H, probability measure (over $X \quad f \ 1; +1 g$), and empirical counterpart b be given. Then the weak PAC-learning rate satisfies ${}_0( )$, and the empirical weak learning rate (b) satisfies (b) ${}_0(b)$.

The following example highlights why ${}_0( )$ can be problematic, even when $L = 0$.

Example 6.3.4. (Nightmare scenario #1.) Suppose $X = (0 ; 1]$, and

$$\Pr[y = +1 \, jx] = \begin{cases} 1 & \text{when } x \in (1=(i+1) ; 1=i] \text{ for some odd integer } i; \\ 0 & \text{when } x \in (1=(i+1) ; 1=i] \text{ for some even (positive) integer } i: \end{cases}$$

Let H consist of threshold functions (decision stumps). Given any integer k, a combination of thresholds may be constructed which is correct on k intervals, and thus $L = 0$ by considering

$k \to 1$. Unfortunately, the norm of these solutions also grows unboundedly, suggesting and $(\ )$ are tiny. Indeed, consider a distribution over X which is uniform on k of the intervals, and zero elsewhere. Any threshold is incorrect on nearly half of these intervals, and by considering $k \to 1$, it follows that $_0(\ ) = 0$.

In precise terms, this nightmare, and suggested sequence of distributions, provide the following property.

Proposition 6.3.5. There exist choices for H and so that $L = 0$, but $_0(\ ) = 0$ and, with any probability 1 and sample size m large enough that $k := \lfloor m^{1/4}/(3\sqrt{2\ln(4/\ )}) \rfloor$ satisfies $k \geq 2$, then $(b) \geq _0(b) \geq (b) - O\left((\ln(4/\ ) + \ln(m))^{1/2}/m^{1/4}\right)$, where $= O(1/k) = O(m^{-1/4}/\ln(4/\ ))$ and the $O(\ )$ only suppresses terms independent of m and .

But something is wrong here | Example 6.3.4 seems quite easy! The reason indicates otherwise is that it simply tries too hard: Example 6.3.4 is easy if giving up on an -fraction of the data is acceptable. This reasoning leads to the relaxation $(\ )$, which, in contrast to Proposition 6.3.5, carries the following guarantee.

Proposition 6.3.6. Let probability over $X \times \{-1, +1\}$ and class H be given.

1. Let loss $` \in L_{lg}$ be given. Then $L = 0$ i $(\ ) > 0$ for all > 0 i $(C) = 0$ for every hard core C.

2. Let any > 0, con dence parameter $\in (0, 1]$, and empirical measure b be given. Then with probability at least 1 ,

$$(b) \geq (\ ) - \sqrt{\frac{1}{2m}\ln\frac{2}{}}:$$

Note that the rst item above also indicates that the two cases considered in this chapter could equivalently be stated in terms of hard cores: the separable case $= 0$ is equivalent to all hard cores having zero measure, and the nonseparable case > 0 is equivalent to all hard cores having positive measure.

In order to prove this result, and also a few other components in the proof of Theorem 6.3.2, the following dual representation of $(\ )$ is used. A similar result was proved by Shalev-Shwartz and Singer (2008, see the quantity AvgMin_k) in the case of measures with nite support and

nite cardinality hypothesis classes; the proof here invokes Sion's Minimax Theorem (Komiya, 1988), which operates in fairly general topological vector spaces.

Lemma 6.3.7. Let probability measure over $X \quad f \quad 1; +1 g$, any H, and any $\quad 2 \ (0; 1]$ be given. Then

$$\begin{aligned} ( ) &= \min_{p2D ( )} \sup \quad (A \ )pd \ : \quad 2 \quad ; k \ k_1 \quad 1 \\ &= \sup \quad \min_{p2D ( )} \quad (A \ )pd \ : \quad 2 \quad ; k \ k_1 \quad 1 \\ &= \min_{p2D ( )} kA^{>} pk_1 \ ; \end{aligned}$$

where $A^{>}$ is the unique adjoint operator to A (cf. Lemma 1.A.5), and

$$kA^{>} pk_1 = \sup \quad yh(x)p(x; y)d \ (x; y) \ : h \ 2 \ H \quad :$$

In order to use this to prove the rst part of Theorem 6.3.2, rst note that whenever $( ) = 0$, there exists a dual element $p \ 2 \ D ( )$ certifying this property, which in turn can be related to the duality structure of L (as presented in Lemma 4.1.1), and gives the result. For the second part of Theorem 6.3.2, similarly the in mum in the de nition of $( )$ can be removed by considering a single bad certi cate $p \ 2 \ D ( )$, and the supremum can be removed by considering a single good $\quad 2$. The certi cate p can be shown to have a simple structure (it emphasizes margin violations for the xed good), and in turn the deviations are easy to control.

6.3.2 Proof Sketch of Theorem 6.3.2

The pieces are in place to establish the nite sample guarantees in Theorem 6.3.2. First, note the following empirical risk guarantee.

Lemma 6.3.8. Let any $` \ 2 \ L_{lg}$, empirical measure b, and H be given. Suppose Figure 1.3 is run with any of the three step size choices for T iterations, let $_t = R^b(H_t)$ denote the classi cation error of H_t, and set $^0_t := _{t \ 1} = _2$ for convenience. Then

$$L^b(A_T) \quad `(0) \quad \sum_{t=1}^{T} \frac{(\quad _1 \ _t \ _1 \ ^0_{t \ 1}(b))^2}{6B_2} :$$

Notice that this result indicates that the convex risk decreases quickly in the presence of

classi cation errors. The proof, sketched as follows, is fairly straightforward. First, standard properties of the line search choices show that $\hat{L}$ drops in round t proportionally to $\|A^\top r\hat{L}(A_{t-1})\|_1$. Considering $r\hat{L}(A_{t-1})$ as a reweighting of b, this expression appears in the dual form of dual form of (b) as presented in Lemma 6.3.7. In order to make the correspondence precise, $r\hat{L}(A_{t-1})$ must be rescaled to unit norm; but, by the Lipschitz property, the rescaling is by at most $1/t-1$! After some algebra, and summing across all iterations, the result follows.

From here, there is little to do. By Lemma 6.3.8, until some iteration has low error, progress is quick. The selection rule (returning $\hat{}$ with minimal classi cation risk) ensures there are no problems if the classi cation risk happens to go back up, and Proposition 6.3.6 allows ( ) to replace (b). As this reasoning provides a direct guarantee on the empirical classi cation risk, standard uniform convergence techniques give the result.

6.4 The Nonseparable Case (L > 0)

When $L > 0$, the essential object will be an optimum to the convex dual of the central optimization problem $\inf L(A\ )$, speci ed as follows. The following slight re nement of the duality result from Lemma 4.1.1 additionally speci es a few properties of the dual optimum.

Proposition 6.4.1. Let loss $\ell \in L_{lg}$ (with tightest Lipschitz parameter $_2$), class H, and probability measure over $X \quad f-1;+1g$ be given. Then

$$\inf \int \ell(A\ )d\ :\ \in \quad = \max \int \ell^*(p) : p \in L^1(\ ); p \in [0;\ _2]\ \text{-a.e.}; \|A^\top p\|_1 = 0\ ;$$

where ℓ^* is the Fenchel conjugate to ℓ, and the adjoint $A^\top$ is as in Lemma 6.3.7 and Lemma 1.A.5. Additionally, the dual optimum p satis es $([p\ ])\ $, where > 0 whenever the optimal value L is positive, and moreover has the explicit form $:= \ell^{*-1}(-L)=2$, where ℓ^{*-1} is the (well-de ned) inverse of ℓ^* along $[0;\ell^{*\prime}(0)]$.

The strategy in the nonseparable case is to exhibit curvature in the objective function (i.e., a lower bound on the second-order expression $m^{-1}\sum_i \ell((A\ )_{x_i;y_i})$), and the dual optimum p will be provide the mechanism; making these statement precise is the topic of this section.

Theorem 6.4.2. Let loss $\ell \in L_{lg} \cap L_{2d}$, binary class $H \in F_{vc}$, probability measure over $X \quad f-1;+1g$ with empirical counterpart b corresponding to a sample of size m, time horizon

t m^a with $a \in (0;1)$, and any con dence $\in (0;1]$ be given. Suppose $L > 0$, and let $p \in L^1(\ )$ denote the dual optimum as in Proposition 6.4.1, with corresponding real number so that $([p\ ])$. De ne the quantities

$$c := \frac{16\grave{}(0)}{\grave{}^0(0)} \max\ 1; \frac{1}{-}\ \max\{1; kpk_1\ g; \qquad B_1 := \frac{}{8}\ \inf_{z \in [\ c;+\ c]} \grave{}^{00}(z);$$

$$R_i := \sqrt{\frac{\grave{}(0)\max\{5; 2B_1=B_2 g}{^2 B_1}};$$

and suppose the sample size is large enough to satisfy $(2=\ ^2)\ln(4=\ )$ and

$$\frac{2(R_t + 2\,c)kpk_1}{m^{1=2}}\ \ 2^p\overline{2V(H)\ln(m+1)} + \sqrt{2\ln(4=\ )}\qquad \frac{c^2}{8}$$

(which happens for all large m since $\lim_{m!1}\ R_t\sqrt{\ln(m)=m} = \lim_{m!1}\sqrt{\ln(m)=m^{1\ a}} = 0$). Then it holds that the above values , c, B_1, and R_i (for $0 < i\ t$) are all positive, and moreover the following statements hold simultaneously with probability at least $1\ $.

1. The nal coe cient vector $_t \in$ satis es

$$L(A\ _t)\ \ \inf_{k\ k_1\ R_{t\ 1}} L(A\ ) + m^{\ a=4} + R_{t\ 1}\ \sqrt{\frac{2}{m}\ln\frac{6}{-}}$$
$$+ \frac{2\ _2 R_t}{m^{1=2}}\ \ 2^p\overline{2V(H)\ln(m+1)} + \grave{}(R_t)\sqrt{2\ln(6=\ )}$$
$$+ \grave{}(0)\ \left(\frac{k\ k_1}{k\ k_1 + m^{\ a=4}=(4B_2R_1)}\right)^{3B_1=(9B_2)}:$$

2. If $H \in F_{ds}(\ ^X)$ (where ^{X} is the marginal of over X), and letting $R_?$ denote the Bayes error rate, there exists $: R ! R$ satisfying $(R(A\ _t)\ R\ _?)\ L\ (A\ _t)\ L$ and $(z) ! 0$ as $z ! 0$. (For instance, when $\grave{} = \grave{}_{log}$, then $(z) = z^2=2$.)

3. The returned coe cients $\hat{}$ satisfy

$$R(H\hat{}\ )\ \ R\ (H\ _t) + 4\ \ R(H\ _t)\sqrt{\frac{(V(H)+1)\ln(2\ em) + \ln(24=\ )}{m^{1\ a}}}$$
$$+ 8\,\frac{(V(H)+1)\ln(2\ em) + \ln(24=\ )}{m^{1\ a}}:$$

This bound is inferior to the guarantee in the separable case; while it is still of the form

$1=m^{1=c}$, the exponent $1=c$ is distribution-dependent. The source of weakness is the optimization guarantee (cf. Lemma 6.4.6), which is brute-forced and should be improvable.

6.4.1 Curvature

Recall that the dual optimum p satisfies $\|A^\top p\|_1$, which implies $\int (A \ )p d \ = 0$ for every 2 (cf. Lemma 1.A.5). To see how this helps locate bad examples and produce curvature, note the rearrangement

$$\int_{[A < 0]} (A \ )p d \ = \int_{[A \ 0]} (A \ )p d \ ;$$

meaning has been reweighted by$p$ so that negative and positive margins are equal (in a sense, p renders every 2 equivalent to random guessing). Since$p$ is fairly well-behaved (it is within $[0; \ _2]$ -a.e. (where $_2$ is the Lipschitz constant for $`$), and is fairly at since $([p \ ])$), then some algebra allows the removal of$p$ from the above display, which yields the statement: if A has many good margins, it also has many bad margins. This constrains the norms of solutions found by the algorithm, and generates curvature in the sense that progress in any direction quickly leads to$L$ increasing.

Of course,p could have instead been directly constructed from the presence of noise, but then the results would not be applicable to cases where itself is noiseless, but$H$ is simply very weak. The following example emphasizes this role of noise, but also shows that the above development overlooked the e ect of sampling.

Example 6.4.3. (Nightmare scenario #2.) Pick any X, (marginal) distribution $_X$ over X, hypothesis class$H \ 2 \ F_{ds}(\ _X)$, and any 2 . De ne the conditional density $Pr[y = +1 \ jx]$ to be 0.9 when $(H \ )_x \ 0$, and 0.1 otherwise when $(H \ )_x < 0$. By this construction, H attains the Bayes error rate (which is 0.1), and every other$H$ does at best this well. Any weighting with favorable convex risk$L(A \ )$ will necessarily have a small norm in consequence of the guaranteed 10% classi cation error.

Unfortunately, nite samples look slightly di erent. Suppose $X = R^d$ and $_X$ is absolutely continuous with respect to Lebesgue measure. With probability 1, a random sample of any size will contain no noise, and span(H) has a perfect predictor $H^{\sim}$ (over the sample); in particular, nothing inhibits the norms of solutions over L^b.

In this example, the good predictor $H^{\sim}$ is potentially very complex, as it is tting noise.

The solution here will be to only control those predictors with small norms; note that this deviation inequality embeds the reweighted average margin expression from eq. (6.2.1).

Lemma 6.4.4. Let probability measure β over $X \times \{-1,+1\}$ with empirical counterpart $\hat{\beta}$, any hypothesis class $H \subseteq F_{vc}$, reweighting $p \in L^1(\beta)$ with $\|A^\top p\|_1 = 0$, and norm bound C be given. Then, with probability at least $1-\delta$, $p \in [0, \|p\|_1]$ $\hat{\beta}$-a.e., and

$$\sup_{\substack{k \\ k_1^2 \le C}} \int (A_\cdot)\, p\, d(\hat{\beta} - \beta) \le \frac{2C\|p\|_1}{m^{1/2}} \left(2\sqrt{2V(H)\ln(m+1)} + \sqrt{2\ln(1/\delta)}\right).$$

Armed with these tools, the structure of the nonseparable problem is as follows. Note that the term B_1 is the aforementioned curvature lower bound, and furthermore the facts $\sum_i \alpha_i = 1$ and $\sum_i \alpha_i^2 < 1$ mean that the step sizes exactly fit the constrained step size regime studied by Zhang and Yu (2005, Equation (4)).

Lemma 6.4.5. Suppose the setting and quantities in the preamble of Theorem 6.4.2; the following statements hold simultaneously with probability at least $1-\delta/2$.

1. Every w with $\|w\| \le R_t + 4c$ and $\hat{L}(A_\cdot w) < 2\ell(0)$ has $m^{-1} \sum_{i=1}^m \ell''((A_\cdot w)_{x_i,y_i}) \ge B_1$.
2. For every choice of step size $\alpha_i \le R_i$ and
$$\alpha_i^2 \le \min\left(\frac{9\|A^\top r\|\, \hat{L}(A_{i-1} w_{i-1}) k_1^2}{4\ell''(0) B_1^2}, \frac{\max\{5, 2B_2/B_1\}(\hat{L}(A_{i-1} w_{i-1}) - \hat{L}(A_i w_i))}{\ell''(0) B_1}\right).$$
3. Let w with $\|w\| \le R_1 \sqrt{t-1} = R_{t-1}$ and $\gamma := \min_{i \in [t-1]} \hat{L}(A_i w_i) - \hat{L}(A w) > 0$ be arbitrary. For every choice of step size,
$$\alpha_i \le \frac{(\hat{L}(A_{i-1} w_{i-1}) - \hat{L}(A w))}{2B_2(\|w\| + R_1\sqrt{i-1})} \quad \text{and} \quad \sum_{i=1}^t \alpha_i \le \frac{\sqrt{t-1}}{4B_2 R_1}.$$

6.4.2 Proof of Theorem 6.4.2

The convergence analysis due to Zhang and Yu (2005) can be adjusted to the present setting (where step and coordinate selection are decoupled), yielding the following guarantee. Note that Lemma 6.4.5 also allows the application of the analysis due to Bartlett and Traskin (2007) (again with decoupling modifications), however this leads to a rate of roughly $O(1/\sqrt{\ln(m)})$.

Lemma 6.4.6. Let $\ell \in L_{lg} \cap L_{2d}$ with Lipschitz gradient parameter B_2, binary class H, time horizon t, and empirical probability measure b be given. Let $\alpha > 0$ be arbitrary, and suppose there exists $c_3 > 0$ with $c_3 \cdot \|A - A'\| \geq |L(A) - L(A')|$ for for all $0 \leq i \leq t$. Then

$$L(A_t) - L(A) \leq L(A_0) - L(A) \leq \frac{k k_1}{k k_1 + \sum_{i=1}^{t} i} \left(\frac{\alpha}{c_3^2/(6B_2)}\right)^{c_3^2/(6B_2)}.$$

From here, there is little to do: the conditions for this rate are met with high probability thanks to Lemma 6.4.5, and the rest is standard uniform convergence.

Appendix 6.A The Family of Dense Classes $F_{ds}(\cdot)$

As the goal of a consistency analysis is to show that the Bayes predictor is approximated arbitrarily nely, necessarily the function class considered by a purportedly consistent algorithm must be very large.

As discussed in Section 6.1, one choice is the class $F_{ds}(\cdot)$ of functions dense according to $L^1(\cdot)$ in the family of bounded measurable functions. A partial survey of density assumptions in other work is as follows.

> Breiman (2000, De nition 1) works with a similar de nition: the relevant metric is $L^2(\cdot)$, and the closure must contain $L^2(\cdot)$, where $\cdot$ is constrained to be continuous with respect to Lebesgue measure. By contrast, the metric for $F_{ds}(\cdot)$ is $L^1(\cdot)$, where $\cdot$ is an arbitrary measure over the Borel σ-algebra, and the closure of the class must contain bounded measurable functions, which are a subspace of $L^1(\cdot)$, which is contained within $L^2(\cdot)$.
>
> Proposition 6.1.5, which will be proved shortly, states that it su ces for $\mathrm{span}(H)$ to contain boxes formed by half-open intervals. This result was stated by Breiman (1999, Proposition 1) with an abbreviated proof for his setting of Lebesgue-continuous measures, thus the present result can be taken as merely proving that result with slightly more generality and verbosity.
>
> The closest assumption and family of results to those here were provided by Zhang (2004, Section 4); while an analog to Proposition 6.1.5 is not shown there, the proofs rely on a form of Lusin's Theorem, which is used in Proposition 6.1.5 as well; indeed, the proofs here owe their existence to those earlier ones by Zhang (2004, Section 4).

Another approach, suggested by Lugosi and Vayatis (2004, Theorem 1 and subsequent remarks), and later used by Bartlett and Traskin (2007, Condition 1) and Schapire and Freund (2012, eq. (12.11)), is to require the weaker condition that

$$\inf\left\{L(A) : \quad 2 \quad g = \inf \int `(\ yf(x))\,d\ (x;y) : f \text{ measurable from } X \text{ to } R \quad ;\right.$$

for a veri cation that this property is indeed weaker, see Lemma 6.A.1. Lugosi and Vayatis (2004, Lemma 1) show that this assumption is satis ed by classes whose convex hull contains indicators of all Borel sets, and thus Lemma 6.A.1 can be considered a simpli cation which su ces to grant consistency with more computationally tractable classes (like decision lists and trees).

As discussed above, the essential property of $F_{ds}(\ )$ is that it implies the weaker condition used by Lugosi and Vayatis (2004, Theorem 1 and subsequent remarks), which in turn is directly needed for the classi cation calibration methods in the consistency proof (cf. Theorem 6.2.2). The Lipschitz condition here is not crucial, and for instance can be removed by adjusting $F_{ds}(\ )$ to require approximants to a function to carry nearly the same uniform bound.

Lemma 6.A.1. Let distribution over $X \quad f \quad 1;+1g$, class $H \ 2 \ F_{ds}(\ ^X)$, and nonnegative Lipschitz convex loss $`$ be given (with Lipschitz constant $_2$). Then

$$\inf\left\{L(A) : \quad 2 \quad g = \inf \int `(\ yf(x))\,d\ (x;y) : f \text{ measurable from } X \text{ to } R \quad :\right.$$

Proof. One direction is immediate, since H de nes a family of measurable functions.

Going the other direction, rst de ne, for any measurable f, a clamping

$$[f\ ]_r(z) := \begin{cases} f(z) & \text{when } f(z) \quad r, \\ r & \text{otherwise} \end{cases} :$$

For any > 0, based on four cases for the structure of $`$, a clamping value $r$ is de ned as follows in order to satisfy, for any $f$ and $z$, $`([f\ ]_r(z)) \quad `(f(z)) + \ .$

If $\lim_{z!1} \quad `(z) = \lim_{z! \ +1} `(z) < 1$, then $`$ is a constant function, and $r = 0$ su ces.

If $\lim_{z!1} \quad `(z) = \lim_{z! \ +1} `(z) = 1$, then $`$ has compact level sets, and in particular an

r exists so that

$$\{z : \ell(z) \le \inf_q \ell(q) + \varepsilon\} \subset \{z : |z| \le r\}.$$

It follows that $\ell([f]_r(z)) \le \ell(f(z))$.

If $\lim_{z \to -1} \ell(z) < 1$ and $\lim_{z \to +1} \ell(z) = 1$, then set

$$r := \inf\{|z| : \ell(z) \le \inf_q \ell(q) + \varepsilon\}.$$

Unlike the preceding two cases, clamping here can increase the value, but not by more than ε.

If $\lim_{z \to -1} \ell(z) = 1$ and $\lim_{z \to +1} \ell(z) < 1$, then this case is handled by the preceding one by considering the reflection $z \mapsto \ell(-z)$.

Consequently, let $\{f_i\}_{i=1}^{\infty}$ be a minimizing sequence for the target infimum above so that

$$\int \ell(yf_i(x))\,d\mu(x,y) \le 2^{-i} + \inf\left\{\int \ell(yf(x))\,d\mu(x,y) : f \text{ measurable from } X \text{ to } \mathbb{R}\right\}.$$

Each f_i might not be bounded, so define $g_i := [f_i]_{r_i}$ where $r_i := 2^{-i}$; by this choice,

$$\begin{aligned}\int \ell(yg_i(x))\,d\mu(x,y) &= \int \ell(y[f_i]_{r_i}(x))\,d\mu(x,y)\\ &\le 2^{-i} + \int \ell(yf_i(x))\,d\mu(x,y)\\ &\le 2^{-i+1} + \inf\left\{\int \ell(yf(x))\,d\mu(x,y) : f \text{ measurable from } X \text{ to } \mathbb{R}\right\}.\end{aligned}$$

Lastly, since $\operatorname{span}(H)$ is dense in the $L^1(\mu_X)$ metric, let $h_i \in \operatorname{span}(H)$ satisfy $\|h_i - g_i\|_1 \le 2^{-i}$; since ℓ is Lipschitz with constant α_2, then

$$\begin{aligned}\int \ell(yh_i(x))\,d\mu(x,y) &\le \int \ell(yh_i(x))\,d\mu(x,y) + \int (\ell(yg_i(x)) - \ell(yh_i(x)))\,d\mu(x,y)\\ &\le \int \ell(yh_i(x))\,d\mu(x,y) + \alpha_2 \int |yg_i(x) - yh_i(x)|\,d\mu(x,y)\\ &\le \int \ell(yh_i(x))\,d\mu(x,y) + \alpha_2 \|g_i - h_i\|_1\\ &\le (2 + \alpha_2)2^{-i} + \inf\left\{\int \ell(yf(x))\,d\mu(x,y) : f \text{ measurable from } X \text{ to } \mathbb{R}\right\},\end{aligned}$$

and the result follows. □

To close, the proof of Proposition 6.1.5, which avoids strong structural assumptions on the measure (for instance, a relationship to Lebesgue measure) via an invocation of Lusin's Theorem.

Proof of Proposition 6.1.5. Let > 0 and bounded measurable g with $\|g\|_u := \sup_x |g(x)| \in (0, 1)$ (when $\|g\|_u$, then $g \in \text{span}(H)$ and the proof is complete). By Lusin's Theorem, there exists compactly-support continuous $h \in L^1(\ )$ which satisfies $([h \neq g]) \quad =(4\|g\|_u)$, and $\|h\|_u \quad \|g\|_u$ (Folland, 1999, Theorem 7.10). Let C denote the compact support of h; continuity over a compact subset of R^d means uniform continuity, and therefore let > 0 be sufficiently small that the bounding box of C may be partitioned into finitely many cubes of side length (products of half-open intervals of length) so that, for any x_1 and x_2 within a single cube, $|h(x_1) \quad h(x_2)| \quad =2$. Now let f be a sum of indicators of these cubes, where each indicator is weighted by $h(x)$ with x being an arbitrary point in the corresponding cube. By construction and since $\text{span}(H)$ contains such cubes, $f \in \text{span}(H)$, and moreover $\|f \quad h\|_1 \quad =2$ since is a probability measure, which provides

$$\|f \quad g\|_1 \quad \|f \quad h\|_1 + \|h \quad g\|_1 \quad =2 + \int_{[h \neq g]} |h \quad g| d \quad =2 + 2\|g\|_u \quad ([h \neq g]) \quad :$$

□

Appendix 6.B Loss Function Classes L_{lg} and L_{2d}

First, note that L_{lg} and L_{2d} contain a few useful things.

Lemma 6.B.1. $\ell_{log} \in L_{lg}$ with parameters $B_2 = 1/4$, $_1 = 1/2$, $_2 = 1$. $\ell_{russ} \in L_{lg}$ with parameters $B_2 = \ _1 = \ _2 = 1$. Lastly, $\exp \in L_{2d}$ and $\ell_{log} \in L_{2d}$.

Proof. For the logistic loss ℓ_{log}, note $0 \quad \sup_x \ell''_{log}(x) \quad 1/4$, thus the mean value theorem grants Lipschitz gradients with parameter $B_2 \quad 1/4$. ℓ_{log}'s Lipschitz parameters are $_1 = 1/2$ and $_2 = 1$.

Since ℓ_{russ} is not twice differentiable, gradient slopes must be checked manually. To start,

note

$$\grave{}'_{\text{russ}}(x) = \begin{cases} 0 & \text{when } x \quad 1; \\ x + 1 & \text{when } x \ 2 \ (\quad 1; 0); \\ 1 & \text{when } x \quad 0; \end{cases}$$

whereby $_1 = \ _2 = 1$. Within each line segment, the gradient slopes are 0, 1, and 0. By manually checking pairs $x < y$ in the first and second, first and third, and second and third intervals, the tightest Lipschitz constant on the gradients is 1.

The containments within L_{2d} are direct. □

Next, note that the Lipschitz gradient properties on $\grave{}$ also cover $L^{\flat}$.

Lemma 6.B.2. Let convex differentiable $\grave{} : R \ ! \ R$ with Lipschitz gradient parameter B_2, any H, and any empirical measure b over $X \ f \quad 1; +1 g$ be given. Then $L^{\flat} \ A$ has Lipschitz gradients with parameter B_2 with respect to norm $L^1(b)$ over the t^{th} level set

$$S_t := \{ \quad 2 \ : \ L^{\flat}(A \ ) \quad L^{\flat}(A \ _{t \ 1}) \}:$$

Proof. Starting from Lemma 1.C.2, for any $; \ ^0 2 S_t$, since $\grave{}'' \quad 0$ (whereby $j\grave{}''j = \grave{}''$), and also using the definition of the level set S_t,

$$\begin{aligned} r \ (L^{\flat} \ A)(\ ) \ r \ (L^{\flat} \ A)(\ ^0) \ _1 & \quad \frac{1}{m} \sum_{i=1}^{m} j\grave{}'((A \ )_{x_i;y_i}) \quad \grave{}'((A \ ^0)_{x_i;y_i})j \\ & \quad \frac{1}{m} \sum_{i=1}^{m} B_2 j(A \ )_{x_i;y_i} \quad (A \ ^0)_{x_i;y_i} j \\ & \quad \frac{1}{m} \sum_{i=1}^{m} B_2 \quad y_i \sum_{h2H} h(x_i)(\ (h) \quad ^0(h) \\ & \quad B_2 k \quad ^0 k_1 \end{aligned}$$

as desired. □

As a final basic result about $\grave{}$, note that the terminology \separable" is at least somewhat justified.

Proposition 6.B.3 (See also Theorem 5.4.2) Suppose $\grave{} : R \ ! \ R_{++}$ is convex with $\lim_{z \ ! \ 1} \grave{}(z) = 0$, and let any H and any probability measure over $X \ f \quad 1; +1 g$ be given. Suppose $\inf \int_R \grave{}(A \ )d \ =$

0.

1. For any > 0, there exists 2 so that $([A \quad 1]) \quad 1 \quad$.
2. With probability 1 over the draw of a sample $f(x_i; y_i)g_{i=1}^m$ (for any $m < 1$), there exists 2 so that $(A \ )_{x_i;y_i} \quad 1$ for every i.
3. In general, there does not exist 2 so that $([A \quad 0]) = 1$ (indeed, Example 6.3.4 provides a counterexample).

Proof. Let > 0 be given, and choose $f_i g_{i=1}^1$ so that $\int `(A \ )d \quad 1=i$. Since $`(A_i) ! \ 0$ -a.e., by Egoro 's theorem there exists S with $(S) \quad 1$ so that $`(A_i) ! \ 0$ uniformly on $S$ (Folland, 1999, Theorem 2.33). But since $` > 0$ everywhere and $\lim_{z! 1} `(z) = 0$ and $`$ is convex, it must be the case that $(A_i) ! \ 1$ uniformly on S, and so there exists i with $A_i \quad 1$ on S, which gives the first result.

For the second result, take any > 0, and choose as granted by the first part. Let b denote the empirical measure over the provided sample; then

$$\Pr[8i \quad (A \ )_{x_i;y_i} \quad 1] = \quad ([A \quad 1])^m \quad (1 \quad)^m :$$

Since > 0 was arbitrary, the second result follows.

For the third result, recall that Example 6.3.4 (whose properties are provided in Proposition 6.3.5) gave an instance where every element of $\operatorname{span} H()$ makes some mistakes. □

Appendix 6.C Duality Properties of

In order to develop $(\ )$, the set $D(\ )$ must first be studied.

Proposition 6.C.1. (Basic properties of $D(\ )$.) Let be an arbitrary probability measure over $X \quad f \quad 1; +1g$, and let $2 [0; 1]$ be arbitrary. The set $D(\ )$ has the following properties.

1. $D(\ )$ is convex.
2. $D(\ )$ is closed in the $L^1(\ )$ topology.
3. If > 0, then $D(\ )$ is closed in the $L^1(\ )$ topology, and also closed in the weak* topology (i.e., the weak topology induced upon $L^1(\ )$ by $L^1(\ )$).

4. $D(\cdot)$ is compact in the weak* topology on $L^1(\cdot)$ (as discussed in the preceding point).

5. $D(\cdot)$ is not guaranteed to be compact in the $L^1(\cdot)$ or $L^1(\cdot)$ topologies; indeed, it is not compact when $= 1=2$, $X = [0;1]$, the marginal distribution X is uniform on $[0;1]$, and the conditional distribution $\Pr(Y = 1 \mid X = x)$ is arbitrary.

Proof. 1. For convexity, let any $2 (0;1)$ and $p_1; p_2 2 D(\cdot)$ be given, and define sets $N_j := p_j^{-1}((1 ;0))$ for $j 2 f1;2g$, where necessarily $(N_j) = 0$. The goal is to show $p := p_1 + (1)p_2 2 D(\cdot)$.

Define $N := N_1 [N_2$ (where again $(N) = 0$). First, for any $(x;y) 2 N^c$,

$$p(x;y) = p_1(x;y) + (1)p_2(x;y) \quad 0 + (1) 0 = 0;$$

whereby it follows that $p \quad 0$ -a.e.. Second,

$$kpk_1 \quad kp_1k_1 + (1)kp_2k_1 \quad 1= ;$$

again using the convention $1=1 = 0$, whereby $kpk_1 \quad 1=$ as desired. Lastly,

$$kpk_1 = \int j p_1 + (1)p_2j = \int_{N^c} (p_1 + (1)p_2) = kp_1k_1 + (1)kp_2k_1 = 1;$$

meaning all conditions are met, and $p 2 D(\cdot)$. Since $; p_1$, and p_2 were arbitrary, it follows that $D(\cdot)$ is convex.

2. For closure within $L^1(\cdot)$, since $L^1(\cdot)$ is a metric space, it is first countable, and thus it suffices to check that any sequence $fp_j g_{j=1}^1$ with $p_j 2 D(\cdot)$ and $p_j ! p 2 L^1(\cdot)$ satisfies $p 2 D(\cdot)$ (Folland, 1999, Proposition 4.6). Given any such sequence $fp_j g_{j=1}^1$, choose a subsequence $fq_i g_{i=1}^1$ so that $q_i ! p$ -a.e. (Folland, 1999, Corollary 2.32).

Let N_p be the (null) set of points for which convergence fails, and additionally, for each i, define $N_i := q_i^{-1}((1 ;0))$; lastly, set $N := N_p [([_i N_i)$, where again $(N) = 0$. Thus for

any $(x;y) \in N^c$,

$$\begin{aligned} p(x;y) &= q_i(x;y) + (p(x;y) - q_i(x;y)) \\ &\leq q_i(x;y) + |p(x;y) - q_i(x;y)| \\ &\leq \liminf_{i\to\infty} q_i(x;y) + |p(x;y) - q_i(x;y)| \\ &= 0; \end{aligned}$$

thus $p \leq 0$ a.e.. Additionally,

$$\|p\|_1 = \int |p| = \int_N p = \int_N p_i + \int_N (p_i - p) = \|p_i\|_1 + \int (p_i - p);$$

whereby

$$\|p\|_1 - \|p_i\|_1 = \int (p_i - p) \leq \|p_i - p\|_1 \to 0;$$

and $\|p\|_1 = 1$ as desired.

For the last property, if $\alpha = 0$, there is nothing to show, thus suppose $\alpha \in (0;1]$, set $P_i := q_i^{-1}((1-\alpha;1])$, and $Z := N_p \cup (\cup_i P_i)$, whereby it follows that

$$\mu(Z) = 0; \quad q_i \to p \text{ over } Z^c; \quad q_i \geq 1-\alpha \text{ over } Z^c.$$

Then, for any $(x;y) \in Z^c$,

$$p(x;y) = q_i(x;y) + (p(x;y) - q_i(x;y)) \leq \limsup_{i\to\infty} q_i(x;y) + |p(x;y) - q_i(x;y)| \leq 1-\alpha;$$

which establishes $\|p\|_1 \leq 1-\alpha$, and thus $p \in D(\alpha)$.

3. Note first that if $\alpha = 0$, then $D(\alpha)$ can contain members which are not elements of $L^1(\mu)$, and thus discussing this set in the $L^1(\mu)$ topology does not make sense. For the remainder of this case, suppose $\alpha > 0$.

   Just as in the case of $L^1(\mu)$, for $L^1(\mu)$ it suffices to let a sequence $\{p_i\}_{i=1}^{\infty} \subset D(\alpha)$ be given with $p_i \to p$ in the $L^1(\mu)$ topology, and to show that $p \in D(\alpha)$. Notice however, since α is

a probability measure, that

$$\int |p_i - p| \, d\mu = \int |p_i - p_1| \, d\mu \leq \int |p_i - p| \, d\mu + \int |p - p_1| \, d\mu = \|p_i - p\|_1 ;$$

meaning $p_i \to p$ in $L^1(\mu)$ as well, which by the preceding case provides that $p \in D(\mu)$ as desired.

Lastly, since $D(\mu)$ is convex and additionally closed according to $L^1(\mu)$, then it is also weak* closed (Rudin, 1973, Theorem 3.12).

4. Again suppose $\varepsilon > 0$, and define

$$B := \{p \in L^1(\mu) : \|p\|_1 \leq 1/\varepsilon\}.$$

By Alaoglu's Theorem (Folland, 1999, Theorem 5.18), B_1 is compact in the weak* topology, thus $B = \varepsilon^{-1} B_1$ is weak*-compact as well. The result follows since $D(\mu)$ is a weak*-closed subset of B, and closed subsets of compact sets are compact (Folland, 1999, Theorem 4.22).

5. Noncompactness can be understood from the fact that norm balls are in general not compact, but an explicit construction is provided for completeness. Since both $L^1(\mu)$ and $L^1(\nu)$ are metric spaces, to prove non-compactness, it suffices to prove $D_{1/2}(\mu)$ is not totally bounded. In particular, a countably infinite subset of $C \subset D_{1/2}(\mu)$ will be constructed satisfying the property $(f, g) \in C \times C$ with $f \neq g$ implies $\|f - g\|_1 = 1/2$ and $\|f - g\|_1 = 2$, which suffices to show that C (and thus $D_{1/2}(\mu)$) is not totally bounded (in either metric) for the following reason. Let S be any finite subset of $L^1(\mu)$ or $L^1(\nu)$. Since C and S have respectively infinite and finite cardinalities, there must exist $h \in S$ which is a closest element in S to two distinct functions $f \neq g$ in C. Let $\|\cdot\|$ denote either norm under consideration, and note that

$$1/2 \leq \|f - g\| \leq \|h - f\| + \|h - g\| \leq 2 \max\{\|h - f\|, \|h - g\|\},$$

which means that one of these two distances is at least $1/4$. Since S was an arbitrary finite set, it follows that there is no finite set of balls of radius $1/8$ which covers C, and thus C and $D_{1/2}(\mu)$ are not totally bounded according to either norm.

The construction is as follows. For every positive integer $i \in Z_{++}$, define the function

$$f_i(x,y) := 2^{i-1} \sum_{j=0}^{2^{i-1}} 1_{x \in [(2j)2^{-i}, (2j+1)2^{-i})}.$$

Define $C := \{f_i : i \in Z_{++}\}$. By construction, $C \subset D_{1/2}(\ )$ (i.e., $\|f_i\|_1 = 1$ and $\|f_i\|_\infty = 2$), and moreover $i \neq j$ implies f_i and f_j disagree on exactly half of their support, which yields $\|f_i - f_j\|_1 = 1/2$. and $\|f_i - f_j\|_\infty = \|f_i\|_\infty = 2$.

□

With the structure of $D(\ )$ established, the basic duality structure of $(\ )$ follows. Note that the value of establishing the weak*-compactness of $D(\ )$ is to grant an application of Sion's minimax Theorem without making any topological assumptions on H (or rather, on the subspace H). Additionally, Lemma 6.3.7 in Section 6.3 is a combination of this result and part of Lemma 1.A.5.

Lemma 6.C.2. Let probability measure $\ $ over $X \times \{-1, +1\}$, any H, and any $\ \in [0,1]$ be given. Then

$$\begin{aligned} (\ ) &= \min_{p \in D(\ )} \sup_{\ } \int (A\ )p\,d\ : \ \in \\ &= \sup_{\ } \min_{p \in D(\ )} \int (A\ )p\,d\ : \ \in \\ &= \min_{p \in D(\ )} \|A^\top p\|_1 ; \end{aligned}$$

where $\|A^\top p\|_1$ is discussed in Lemma 1.A.5.

Proof of Lemma 6.C.2. Before applying the duality result, it must be established that the various in ma are attained. To start, consider the nal expression $\inf_{p \in D(\ )} \|A^\top\|_1$, and let $\{p_i\}_{i=1}^\infty$ with $p_i \in D(\ )$ be a minimizing sequence to the in mum. Since Proposition 6.C.1 establishes that $D(\ )$ is weak*-compact, there is a subsequence $\{q_i\}_{i=1}^\infty$ which weak*-converges to some $q \in D$ (Folland, 1999, Theorem 4.29). But Lemma 1.A.5 established that $p \mapsto \|A^\top p\|_1$ is weak* lower semi-continuous, and since it is nite over $L^1(\ )$, it is therefore weak* continuous, and therefore the limit point q attains the in mum. Furthermore, Lemma 1.A.5 provides that $\|A^\top p\|_1$ is the same as the rst in mand, whereby both expressions attain their minimizers and are equal.

The middle expression is the easiest; once again constructing a weak*-convergent sequence $q_i \to q$ with $q_i \in D(\ )$, the definition of weak*-convergence explicitly grants $\int q_i f \to \int q f$ for every $f \in L^1(\ )$, and since $A \in L^1(\ )$ is held fixed within this inner expression, it follows that q attains the infimum.

What remains is to swap minimization and maximization. This in turn follows by Sion's minimax theorem (Komiya, 1988); to verify this application, note that $(p, \ ) \mapsto \int p$ is linear and continuous in both parameters (indeed, this is by construction, since $L^1(\ )$ is isometrically isomorphic to the topological dual to $L^1(\ )$, and the weak* topology over $L^1(\ )$ ensures that this integral relation is continuous for every $\in$), also that is a topological vector space, and lastly that $D(\ )$ is a convex compact subset of a topological vector space (namely, the weak* topology, and not the $L^1(\ )$ topology, where $D(\ )$ is not necessarily compact as per Proposition 6.C.1). □

Appendix 6.D Reweighted Margin Deviations (with p Fixed)

Lemma 6.D.1. Let probability measure over $X \times \{-1, +1\}$ with empirical counterpart b, any hypothesis class $H \in F_{vc}$, reweighting $p \in L^1(\ )$, and norm bound C be given. Then, with probability at least $1 -$, $p \in [0, \|p\|_1]$ b-a.e., and

$$\sup_{\|\|_1 \le C} \left(\int (A\ )p\,db - \int (A\ )p\,d \right) \le \frac{2C\|p\|_1}{m^{1/2}} \left(2\sqrt{2V(H)\ln(m+1)} + \sqrt{2\ln(1/\ )} \right).$$

Proof of Lemma 6.D.1. First, define a simplified reweighting $p'(x,y) := p(x,y)\mathbf{1}[|p(x,y)| \le \|p\|_1]$; by the definition of $\|\|_1$, then $p' = p$ -a.e., and thus, with probability 1, any finite sample of any size has p' and p agreeing. The proof will work with p', which satisfies $\sup_{x,y} |p'(x,y)| \le \|p\|_1$, and then close by discarding a measure zero set and thus relating to p.

The main part of the proof is an almost standard application of Rademacher complexity techniques for voted classifiers (Boucheron et al., 2005a, Theorem 4.1 and its proof, which controls for a surrogate loss and not just the classification loss); the only modification will be to work with a loss function which is sensitive to each example in the sample $S = \{(x_i, y_i)\}_{i=1}^m$, which will require a slightly refined Lipschitz contraction principle for Rademacher complexities (Shalev-Shwartz, 2009, Section 22.2, Lemma 15).

Speci cally, de ne the loss

$$((A)_{x;y}) := \begin{cases} Cp^0(x;y) & \text{when } (A)_{x;y} \quad C; \\ (A)_{x;y}\, p^0(x;y) & \text{when } j(A)_{x;y}j \quad C; \\ +Cp^0(x;y) & \text{when } (A)_{x;y} \quad C: \end{cases}$$

Since $\sup_{h;x} jh(x)j \quad 1$ and $k \quad k_1 \quad C$, it follows that $j(A)_{x;y}j \quad C$, and thus the extremal cases are never encountered, meaning

$$((A)_{x;y}) = (A)_{x;y}\, p^0(x;y);$$

and by construction is Lipschitz with parameter kpk_1 (as a function of (A)) and (A) has uniform bound $Ckpk_1$.

As such, letting R denote Rademacher complexity, by the Lipschitz contraction principle for per-coordinate losses (Shalev-Shwartz, 2009, Section 22.2, Lemma 15), behavior of Rademacher complexity on convex hulls (Boucheron et al., 2005a, Theorem 3.3), and relationship between Rademacher complexity and VC dimension (Boucheron et al., 2005a, See the display after eq. (7)),

$$R(\quad (A)) \quad k\ pk_1\ R((A)) \quad k\ pk_1\ CR(H) \quad k\ pk_1\ C \sqrt{\frac{2V(H)\ln(m+1)}{m}}:$$

This handling of a per-coordinate Lipschitz loss may be inserted into a standard deviation bound for uniformly bounded Lipschitz losses (Boucheron et al., 2005a, Theorem 4.1 and its proof) | albeit with an extra factor two to control deviations in both directions | and it follows, with probability at least $1 \quad$, that

$$\sup_{k\ k_1\ C} \int (A)p^0 db \quad \int (A)p^0 d \quad \frac{2Ckpk_1}{m^{1=2}} \quad 2\sqrt{2V(H)\ln(m+1)} + \sqrt{2\ln(1=)} \quad :$$

To complete the proof, recall that $p^0 = p$ -a.e., and a measure zero event was discarded, whereby $p^0 = p$ b-a.e. as well. □

Appendix 6.E Deferred Material from Section 6.3

6.E.1 Deviations of ( )

This subsection establishes the following one-sided deviation bound on ( ).

Lemma 6.E.1. Let any H, any $\in (0;1]$, any con dence parameter $\in (0;1]$, and any probability measure with empirical counterpart b be given. Then with probability at least 1 ,

$$(b) \qquad ( ) \quad \frac{1}{} \sqrt{\frac{1}{2m} \ln \frac{2}{}} :$$

The di culty in the analysis is that the de nition of ( ) involves an in mum over $p \in D( )$ and a supremum over $\in$ with $k \quad k_1 \quad 1$. The proof strategy employed here is to consider a single good choice for, and to consider the e ect on deviations asp varies. These deviations do not appear to be amenable to the usual approach, as $D( )$ is massive: it is in general not compact in the relevant metric topologies (cf. Proposition 6.C.1), and does not obviously possess other structure granting a uniform convergence result. The approach here is to instead identify that the dual optimum has very simple structure, and moreover this structure is robust to sampling.

Considering again the de nition of ( ), while it is true that $p \in D( )$ is de ned over a potentially massive space, when placed in the expression $\int (A)p\,d$, all that matters is the behavior of p for each value of A , which ranges over $[1;+1]$. That is to say, p is really reweighting the univariate margin distribution of A , and the best it can do is emphasize bad margins. In particular, the following lemma proves basic properties of an idealized univariate distillation of this scenario.

Lemma 6.E.2. Let a probability measure supported on $[1;+1]$ and some $\in (0;1]$ be given. Correspondingly de ne

$$\begin{aligned} S &:= \{c \in [1;+1] : ((1 \quad ;c)) \quad \}; \\ c &:= \sup S ; \\ I &:= (1 \quad ;c); \\ p(r) &:= \frac{1}{} 1[r \in I] + \frac{(\quad (I))}{(\{c\})} 1[r = c] \quad ; \end{aligned}$$

with the convention $0/0 = 0$ in the definition of ρ. These objects have the following properties.

1. S is the closed interval $[-1, c]$.
2. (I) , and $(I \cup \{c\})$.
3. $\|\rho\|_1 = 1$ and $\|\rho\|_1 - 1 = \varepsilon$.
4. The optimization problem

$$\inf \int r\rho(r)\,d\mu(r) : \rho \in L^1(\mu), \|\rho\|_1 = 1, \rho \in [0, 1/\varepsilon] \ \mu\text{-a.e.}$$

is minimized at ρ.

Proof. First note that $-1 \in S$, since μ is supported on $[-1, +1]$ and thus $\mu((-1, -1)) = 0$.

Next, S is an interval, since if $-1 \le c_1 \le c_2$ and $c_2 \in S$, then $\mu((-1, c_1)) \ge \mu((-1, c_2))$ and thus $c_1 \in S$.

To show that S is indeed a closed interval, consider any increasing sequence $\{c_i\}_{i=1}^{\infty}$ with $c_i \in S$, thus $c_i \to c$ for some $c \in [-1, +1]$ since $[-1, +1]$ is compact and the sequence is increasing. Then

$$(-1, c) = \bigcup_{i=1}^{\infty} (-1, c_i),$$

and thus, by continuity of measures Folland (1999, Theorem 1.8),

$$\mu((-1, c)) = \mu\left(\bigcup_{i=1}^{\infty} (-1, c_i)\right) = \lim_{i \to \infty} \mu((-1, c_i)) \ge \limsup_{i \to \infty} \mu((-1, c_i)) ,$$

meaning $c \in S$ and S is closed.

Since S is a closed interval, then $c = \sup S \in S$, and it follows by the preceding properties that $S = [-1, c]$.

By definition, for every $c \in S$, it holds that $\mu((-1, c)) < \infty$, thus $c \in S$ implies that (I) .

Next, for every positive integer $i \in \mathbb{Z}_{++}$, it holds by definition of c that $c + 1/i \notin S$, and thus, again by continuity of measures Folland (1999, Theorem 1.8),

$$\mu([-1, c]) = \mu\left(\bigcap_{i=1}^{\infty} [-1, c + 1/i]\right) \le \liminf_{i \to \infty} \mu([-1, c + 1/i]) :$$

For the norms of p (which is a simple function over the Borel σ-algebra), notice that

$$\|p\|_1 = \frac{1}{\alpha}\,\nu(I) + \frac{\left(\alpha - \nu(I)\right)}{\left(f(c) - g(c)\right)}\left(f(c) - g(c)\right) = 1.$$

Moreover, $p = 1/\alpha$ on $(-1, c)$, and $p = 0$ on $(c, 1)$; to show $\|p\|_1 \leq 1$, the behavior of p on c is all that needs to be checked. Since $\nu((-1, c]) \leq \alpha$, then

$$\nu(I) = \nu((-1, c]) + \nu(f(c) - g(c)) \leq \alpha + \nu(f(c) - g(c)),$$

so $p(c) \leq 1/\alpha$. Additionally $\nu(I) \geq 0$ implies $p(c) \geq 0$, and thus $\|p\|_1 \leq 1/\alpha$ as desired.

Lastly, for the minimization problem, consider any feasible p (meaning $\|p\|_1 = 1$ and $\|p\|_\infty \leq 1/\alpha$) with $\|p - p^*\|_1 > 0$. But since p^* is as large as possible along g, it follows that $p < p^*$ for a positive measure subset of $\mathcal{X}$, and $p > p^*$ for a positive measure subset of $(c, 1)$. Consequently $\int r\,p(r)\,d\mu(r) < \int r\,p^*(r)\,d\mu(r)$. Since p^* was arbitrary, it follows that p^* is a minimal choice. □

The task now is to map the optimization over $D(\cdot)$ down to this idealized univariate search problem. Temporarily adopting notation from probability theory, a first step in this direction would be to write

$$\int_{\mathcal{X}} \nu(A)\,dp = E_p(\nu(A)) = E_p(E(\nu(A) \mid \nu(A) = r)),$$

where the latter notation signifies a conditional expectation with respect the σ-algebra generated by events such that $\nu(A)$ falls in some Borel subset of $\mathbb{R}$ (recall that all σ-algebras here are Borel). In some circumstances, the function $E(\nu(A) \mid \nu(A) = r)$ can be converted into integration over a function that takes r as input, which would directly allow conversion to the above univariate idealization; these techniques generally require assumptions on $\mathcal{X} = \{-1, +1\}$ which would rather be avoided here (Durrett, 2010, Section 5.1.3, regular conditional probabilities). As such, the following result exhibits the desired correspondence manually, albeit keeping the above idea in mind.

Lemma 6.E.3. Let any $\alpha \in (0, 1]$, any probability measure μ over $\mathcal{X} = \{-1, +1\}$, any H, and any $\nu \in \mathbb{R}$ with $\|\nu\|_1 \leq 1$ be given. Define a probability measure μ over $\mathbb{R}$ as the pushforward of

through A, meaning, for any Borel subset S of $\mathbb{R}$,

$$\nu(S) := \mu((A\cdot)^{-1}(S)).$$

Then ν is supported on $[-1;+1]$, and moreover the function $p(x;y) := p((A\cdot)_{x;y})$, where p is as defined in Lemma 6.E.2, is a (feasible) minimizer to the optimization problem

$$\inf \left\{ \int (A\cdot)\, p\, d\mu : p \in D(\mu) \right\}.$$

Proof. Since $\|k\|_1 \leq 1$ and A is a continuous linear operator with unit norm (cf. Lemma 1.A.1, or recall the definition of A and the property $\sup_{x;h} |h(x)| \leq 1$), then $|(A\cdot)_{x;y}| \leq 1$, and thus $(x;y) \mapsto (A\cdot)_{x;y}$ maps $X \times \{-1;+1\}$ to $[-1;+1]$, and so the corresponding pushforward measure ν is supported on $[-1;+1]$. Therefore Lemma 6.E.2 provides the structure of $p \in L^1(\nu)$ attaining the minimum in

$$\inf \left\{ \int r\, p(r)\, d\nu(r) : p \in L^1(\nu); \|p\|_1 = 1; p \in [0;1/\tau] \ \nu\text{-a.e.} \right\}.$$

Setting $p := p_\nu(A\cdot)$ as in the statement, by the above optimality guarantee and by properties of pushforward measures (Resnick, 1999, Theorem 5.5.1),

$$\begin{aligned}
&\inf \left\{ \int r\, p(r)\, d\nu(r) : p \in L^1(\nu); \|p\|_1 = 1; p \in [0;1/\tau] \ \nu\text{-a.e.} \right\} \qquad (6.E.4)\\
&= \int r\, p_\nu(r)\, d\nu(r)\\
&= \int (A\cdot)\, p_\nu(A\cdot)\, d\mu\\
&\inf \left\{ \int (A\cdot)\, p\, d\mu : p \in D(\mu) \right\}.
\end{aligned}$$

Now let $\varepsilon > 0$ and $p \in D(\mu)$ be arbitrary. A corresponding element q with $\|q\|_1 = \|p\|_1$ and $\|q\|_1 \leq \|p\|_1$ will be constructed as follows in order to satisfy

$$\int (A\cdot)\, p\, d\mu \leq \int r\, q(r)\, d\nu(r) + \varepsilon.$$

Cover $[-1;+1]$ with at most $1 + \lceil 1/\varepsilon \rceil$ disjoint half-open intervals $\{I_i\}_{i=1}^k$ of the form $[-1 +$

$i \ 0;\ 1 + (i+1)\ ^0)$ where i is a nonnegative integer and $^0 := = (1 + d1= e)$. De ne

$$q(r) := \sum_{i=1}^{k} (I_i)\ ^1 1[r\ 2\ I_i] \int_{[A\ 2 I_i]} pd\ ;$$

with the convention $0=0 = 0$ (i.e., $q(I_i) = 0$ when $(I_i) = 0$). By this choice, $kqk_1 \quad k\ pk_1$, and

$$kqk_1 = \int q(r)d\ (r) = \sum_{i=1}^{k} \int_{[A\ 2 I_i]} pd\ = kpk_1:$$

More importantly, using Fubini's Theorem to interchange the integrals over and ,

$$\begin{aligned}
&\int (A\ )pd \int rq(r)d\ (r) \\
&\quad \sum_{i=1}^{k} \int_{[A\ 2 I_i]} (A\ )pd \int_{I_i} r\ (I_i)\ ^1 \int_{[A\ 2 I_i]} pd\ d\ (r) \\
&\quad \sum_{i=1}^{k} \int_{[A\ 2 I_i]} (A\ )p\ p \int_{I_i} r\ (I_i)\ ^1 d\ (r)\ d \\
&\quad \sum_{i=1}^{k} \int_{[A\ 2 I_i]} jpjj\ ^0 j\ :
\end{aligned}$$

Since and p were arbitrary,

$$\begin{aligned}
&\inf \int (A\ )pd\ : p\ 2\ D\ (\ ) \\
&\quad \inf \int rp(r)d\ (r) : p\ 2\ L^1\ (\ ); kpk_1 = 1; p\ 2\ [0; 1=]\ \text{-a.e.}\ ;
\end{aligned}$$

which combined with the inequalities starting with eq. (6.E.4) provides that p is indeed a minimizer. □

With these tools in place, the proof of Lemma 6.E.1 follows.

Proof of Lemma 6.E.1. Consider the form of $(\ )$ provided by Lemma 6.C.2, whereby the supremum over 2 is on the outside. Let > 0 be arbitrary, choose 2 which is within > 0 of achieving the supremum, and let$p$ be an optimal dual element as provided by Lemma 6.E.3, together meaning

$$(\ ) \quad + \inf_{p2D\ (\ )} \int (A\ )pd\ = \quad + \int (A\ )p\ d\ : \qquad (6.E.5)$$

Now consider the behavior of$p$ over b. By construction, $kp\ k_{L^1\ (b)} \quad 1= $, however

$kp\ k_{L^1(b)}$ is a random variable; but by Hoe ding's inequality, with probability at least 1 ,

$$kp\ k_{L^1(b)} \quad k\ p\ k_{L^1(\ )} = kp\ k_{L^1(b)} \quad 1 \quad \frac{1}{} \sqrt{\frac{1}{2m} \ln \frac{2}{}};$$

henceforth discard this failure event.

Next instantiate another dual optimum p^b via Lemma 6.E.3, but now over the empirical measure$b$; since 2 is primal feasible in the de nition of (b), and again using the form from Lemma 6.C.2 with the supremum on the outside, it follows that

$$(b) \qquad \int (A\ )p^b db \tag{6.E.6}$$

Now recall the exact form of p and p^b as provided by Lemma 6.E.3 (and more speci cally Lemma 6.E.2), which are both exactly $1=$ up to some point, within $[0; 1=]$ at that point (potentially distinct for p^b and p), and zero thereafter; if $kp\ k_{L^1(b)} \quad 1$, then

$$\int jp \quad p^b jdb = \int p\ db \quad \int p^b db = \int p\ db \quad 1;$$

whereas$kp\ k_{L^1(b)} \quad 1$ implies

$$\int jp \quad p^b jdb = \int p^b db \quad \int p\ db = 1 \quad \int p\ db$$

In either case, using as usual the fact$\sup_{x;y} j(A\ )_{x;y} j \ k \ k_1 \quad 1$, and additionally the controls on $kp^b k_{L^1(b)}$ from above,

$$\int (A\ )p^b db \quad \int (A\ )p\ db \quad \int p^b db \quad \int p\ db$$
$$\int p^b \quad p \quad db$$
$$\frac{1}{} \sqrt{\frac{1}{2m} \ln \frac{2}{}} :$$

Combining this with eqs. (6.E.5) and (6.E.6),

$$(b) \quad \int (A \ )p^{b}db$$
$$\int (A \ )p \ db \quad \frac{1}{} \quad \frac{1}{2m} \ln \quad \frac{2}{}$$
$$( ) \quad \frac{1}{} \quad \frac{1}{2m} \ln \quad \frac{2}{} :$$

Since $>$ 0 was arbitrary, the result follows. □

6.E.2 Other Results

Lemma 6.E.7. Let be a probability measure on X f 1; +1 g. If 0 $_1$ $_2$ 1, then 0 $_1$( ) $_2$( ) 1.

Proof. Let 0 $_1$ $_2$ 1 be given; then $D_{_1}( ) \ D_{_2}( )$ by de nition, and thus $_1$( ) $_2$( ). Next, $_1$( ) 0 follows by Lemma 6.C.2 since $kA^{>} pk_1$ 0, or by considering the e ect of the primal player choosing = 0 2 . For the upper bound, since $\sup_{h;x} jh(x)j$ 1, then $_2$( ) $\sup\{k \ k_1 : \ 2 \ ; k \ k_1 \ 1g$ 1: □

Proof of Proposition 6.3.3. Since every $p \ 2 \ L^1( )$ with $kpk_1 = 1$ and p 0 -a.e. de nes a probability measure pd (ignoring a -null set which does not a ect that value of integration with respect to), and since $k \ e_h k_1 = 1$ for every $h \ 2 \ H$,

$$\inf \ \sup_{h2H} \left(\int yh(x)d \ (x;y) : \ \text{is a Borel probability measure over} \ X \ f \ 1; +1 g \right)$$
$$\inf \ \sup_{k \ k_1 \ 1} \left(\int (A \ )_{x;y} p(x;y)d \ (x;y) : p \ 2 \ L^1( ); kpk_1 = 1 ; p \ 0 \text{ -a.e.} \right)$$
$$= \ _0( ):$$

For ( ) $_0$( ) with a discrete measure over a nite set, the proof is as above (indeed with a tiny re nement, since in this case both ( ) and $_0$( ) consider the same set of weightings over). □

Proof of Proposition 6.3.5. For convenience, de ne $I_i := (1=(i+1) ; i]$. This proof will proceed by establishing, for every k, a bounded weighting p_k, which will establish an upper bound on ( )

for some > 0 which is a function of k. The result will then follow for $_0( )$ by the monotonicity of $( )$ as a function of (cf. Lemma 6.E.7), and result for $_0(b)$ will use deviation bounds on $( )$ and again the monotonicity property.

De ne p_k to be positive over intervals I_i with $1 \quad i \quad 2k$, and zero elsewhere as follows. For any $x \; 2 \; I_i$, $p_k(x;y) = i(i+1)=(2k)$. By this choice,

$$\int_{I_i} p(x;y)\,d\,(x;y) = \frac{i(i+1)}{2k}\left(\frac{1}{i} - \frac{1}{i+1}\right) = \frac{1}{2k};$$

It follows that $kpk_1 = 1$, $kpk_1 = 2k+1$, and p_k makes look like the uniform distribution over k consecutive intervals.

Now consider any hypothesis $h \; 2 \; H$, with some threshold r. If r lies outside this set of intervals, then h is equally correct and incorrect, thus $\int yh(x)p(x;y)d\,(x;y) = 0$. Otherwise, suppose there are a intervals before the threshold, and b intervals after it; h must be incorrect on at least $a=2 \quad 1$ of the left intervals, and $b=2 \quad 1$ of the right intervals; since $2k \quad 1 \quad a+b \quad 2k$ (this proof is charitable), thus at least $k \quad 2$ intervals are predicted completely incorrectly. Consequently,

$$\int yh(x)p(x;y)d\,(x;y) \quad \frac{k+2}{2k} \quad \frac{k \quad 2}{2k} = \frac{2}{k}:$$

Thus the form of $( )$ from Lemma 6.C.2 provides $_{1=2k}( ) \quad 2=k$, and so $_0( ) = 0$ by monotonicity (Lemma 6.E.7), and $\quad 0$ by Proposition 6.3.3.

The remainder of the proof considers nite sample e ects. Let > 0, and a sample of size $m \quad 2$ be given. Choose integer $k := bm^{1=4}=(3\sqrt{2\ln(4=)})c$ (where the lower bound on m provides $k \quad 1$), and consider the behavior of density p_k, de ned as above. Note rstly that $kp_k k_1 \quad 2k+1 \quad 3k \quad m^{1=4}=\sqrt{2\ln(4=)}$. Next, with probability at least $1 \quad =2$, Hoe ding's bound grants

$$\int_{\quad} pdb \quad \int_{\quad} pd \quad k\,pk_1 \sqrt{\frac{1}{2m}\ln\frac{4}{}} \quad \frac{1}{2m^{1=4}} \quad \frac{1}{2}:$$

Now de ne $p_k^0 := p_k=(\int p_k db)$, which means $\int p_k^0 db = 1$, and furthermore $kp_k^0 k_1 \quad 2kp_k k_1$.

Since H has VC dimension $V(H) = 2$, Lemma 6.D.1 grants, with probability at least $1 - \delta/2$,

$$\sup_{h \in H} \left| \int y h(x) p_k^0(x,y) \, db(x,y) - \int y h(x) p_k^0(x,y) \, d\pi(x,y) \right| \leq \frac{2k p_k^0 k_1}{\sqrt{m}} \left(4 \sqrt{\ln(m+1)} + \sqrt{2 \ln(2/\delta)} \right).$$

Now set $\varepsilon := 1/k p_k^0 k_1 \sqrt{2 \ln(4/\delta)}/(2m^{1/4})$. Then $p_k^0 \in D_\varepsilon(b)$, and the above computations provide

$$(b) \quad \frac{16(\sqrt{\ln(m+1)} + 1)}{m^{1/4}} + \frac{2}{k} \leq \frac{16(\sqrt{\ln(m+1)} + 1)}{m^{1/4}} + \frac{2}{m^{1/4}/(3\sqrt{2 \ln(4/\delta)}) - 1};$$

and lastly Lemma 6.E.7 and Proposition 6.3.3 grant $(b) \leq \varepsilon_0(b) \leq \varepsilon(b)$. □

Proof of Proposition 6.3.6. 1. This proof will start by proving $L > 0$ iff $\varepsilon(C) > 0$ for some hard core C iff $\varepsilon(\pi) > 0$ for some $\pi > 0$. The result then follows by negating each of the three statements, and then simplifying them with the facts $L \geq 0$, $\varepsilon(C) \geq 0$, and $\varepsilon(\pi) \geq 0$ for all $\pi \geq 0$. To ease notation, let $S_h(H, \pi)$ denote the set of hard cores.

($L > 0 \implies \exists C \in S_h(H, \pi) \; \varepsilon(C) > 0$.) If $L > 0$, then by Lemma 4.1.1 there exists a dual element p which is dual feasible, meaning $p \in L^1(\pi)$ and $\|A^\top p\|_1 = 0$ and $\pi([p < 0]) = 0$, and moreover p satisfies $\int \ell(p) d\pi < 0$, which means $\pi([p > 0]) > 0$ since $\ell(0) = 0$ by Lemma 1.B.1. As such, simplifying $\|A^\top p\|_1 = 0$ via Lemma 1.A.5, $p \in D(H, \pi)$, and thus $[p > 0] \in S_D(H, \pi)$, and any hard core C satisfies $\varepsilon(C) \geq \pi([p > 0]) > 0$ as desired.

($(\exists C \in S_h(H, \pi) \; \varepsilon(C) > 0) \implies \exists \varepsilon > 0 \; \varepsilon(\pi) = 0$.) Let $p \in S_D(H, \pi)$ be given so that $[p > 0] = C$ for any hard core C (they all have the same measure). Note that $\|p\|_1 > 0$, and define $\bar{p} := p/\|p\|_1$, which satisfies $\|\bar{p}\|_1 = 1$, but still $\bar{p} \in L^1(\pi)$ and $\|A^\top \bar{p}\|_1 = 0$. Since π is a probability measure and $\|\bar{p}\|_1 = 1$, it must hold that $\|\bar{p}\|_1 \geq 1$, and thus the choice $\varepsilon := 1/\|\bar{p}\|_1$ satisfies $\varepsilon \in (0, 1]$. More importantly, since $\bar{p} \in D(\pi)$, it follows that $\varepsilon(\pi) = 0$ as desired.

($(\exists \varepsilon > 0 \; \varepsilon(\pi) = 0) \implies L > 0$.) Let $\varepsilon > 0$ with $\varepsilon(\pi) = 0$ be given; then attainment in the duality formula in Lemma 6.C.2 provides the existence of $p \in L^1(\pi)$ with $\|p\|_1 = 1$ and $\|A^\top p\|_1 = 0$. By Lemma 1.B.1, there exists $c > 0$ so that ℓ is strictly negative along $(0, c)$. Consequently, $\bar{p} := cp/\|p\|_1$ satisfies $\|\bar{p}\|_1 > 0$, and $\bar{p} \in (0, c)$ π-a.e., thus $\ell(\bar{p}) < 0$ π-a.e.,

and also $\|A^\top p\|_1 = 0$; together, it follows that $L \int_{\ell}^{R} \ell'(p)\,d > 0$ as desired.

2. This result is the same as Lemma 6.E.1.

□

Proof of Lemma 6.3.7. This result is the combination of Lemma 6.C.2 and Lemma 1.A.5. □

6.E.3 Optimization Guarantees

Note that the following proof does not overtly use convexity; convexity however is used both algorithmically by the line searches (otherwise they are not efficient), and for their guarantees (cf. Lemmas 1.C.3 and 1.C.7).

Proof of Lemma 6.3.8. Consider any $0 \quad t \quad T \quad 1$. Since $\ell \; 2 \; L_{lg}$,

$$\|\nabla \hat{L}(A_{\quad t})\|_1 \quad \frac{1}{m} \sum_{\substack{i \; 2 \; [m] \\ (A_{\quad t})_{x_i;y_i} \quad 0}} \ell'((A_{\quad t})_{x_i;y_i}) \quad {}_{t \quad 1}.$$

Combining this with the fact that $\ell' 2 [0; \; _2]$, the vector $p_t := \nabla \hat{L}(A_{\quad t})/\|\nabla \hat{L}(A_{\quad t})\|_1$ satisfies $\|p_t\|_1 = 1$ and

$$\|p_t\|_1 \quad \frac{2}{\|\nabla \hat{L}(A_{\quad t})\|_1} \quad \frac{2}{{}_{t \quad 1}} = \frac{1}{{}^{0}_{t}};$$

where ${}^{0}_{t}$ is as provided in the statement (and ${}^{0}_{t} \quad 1$ since ${}_{t} \quad 1$ and ${}_{1} \quad {}_{2}$). Recalling the dual form ${}^{0}_{t}(b) = \min\{\|A^\top p\|_1 : p \; 2 \; D_{\;{}^{0}_{t}}(b)\}$ from Lemma 6.C.2, and noting that $p_t \; 2 \; D_{\;{}^{0}_{t}}(b)$,

$$\|A^\top \nabla \hat{L}(A_{\quad t})\|_1 = \|\nabla \hat{L}(A_{\quad t})\|_1 \|A^\top p_t\|_1 \quad {}_{t \quad 1} \; {}^{0}_{t}(b).$$

Plugging this into the single-step guarantees from the three line search choices (cf. Lemmas 1.C.3 and 1.C.7 and the Lipschitz gradient guarantee from Lemma 6.B.2),

$$\begin{aligned} \hat{L}(A_{\quad t}) \quad \hat{L}(A_{\quad t \quad 1}) \quad & \frac{{}^{2}\|A^\top \nabla \hat{L}(A_{\quad t \quad 1})\|_1^2}{6B_2} \\ \hat{L}(A_{\quad t \quad 1}) \quad & \frac{(\quad 1 \; t \quad 1 \; {}^{0}_{t \quad 1}(b))^2}{6B_2}. \end{aligned}$$

The desired result comes by summing across all iterations and noting $\hat{L}(A_{\quad 0}) = \hat{L}(0) = \ell(0)$. □

6.E.4 Statistical Guarantees

Proof of Theorem 6.3.2. The first step of the proof is to show $\hat{R}(\hat{H}) < \frac{1}{2}$. Thus consider the case that every iteration has $\hat{R}(H_t) > \frac{1}{2}$; by the monotonicity of $\alpha(\cdot)$ (cf. Lemma 6.E.7), positivity of $\alpha(\cdot)$ (cf. Proposition 6.3.6), together with the bound on m (and the deviations on $\alpha(\cdot)$ in Proposition 6.3.6), with probability at least $1 - \delta/2$,

$$\gamma_t'(b) \geq \gamma(b) \geq \gamma(\gamma) - \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}} \geq \frac{\gamma(\gamma)}{2} > 0,$$

where the last equality is also by Proposition 6.3.6. Thus, by Lemma 6.3.8, and the monotonicity of $\ell(\cdot)$ in α, and the second lower bound on m (and thus on m^a),

$$\begin{aligned}
\hat{L}(A_T) &\leq \ell(0) - \sum_{t=1}^{T} \frac{(\alpha_{t-1}\gamma_t(b))^2}{6B^2} \\
&\leq \ell(0) - \frac{m^a(\alpha_{T-1}\gamma(\gamma))^2}{24B^2} \\
&< 0,
\end{aligned}$$

a contradiction since ℓ is nonnegative, and moreover positive on regions where it makes mistakes, therefore the above indicates $\hat{R}(A_T) = 0 < \frac{1}{2}$. As such, thanks to the final step of Figure 1.3 picking out the iterate with lowest classification error, $\hat{R}(\hat{H}) = 0$.

What remains is to establish a deviation inequality. Let $\mathcal{H}_t$ denote the hypothesis class used by predictor f_t (i.e., $\mathcal{H}_t = \{\sum_{i=1}^{t} c_i h_i : c_i \in \mathbb{R}, h_i \in \mathcal{H}\}$), and let $S_{\mathcal{H}_t}(m)$ denote the corresponding shatter coefficient when $\mathcal{H}_t$ is applied to the sample of size m (Boucheron et al., 2005a, Section 3). It follows (Schapire and Freund, 2012, Lemma 4.5) that

$$S_{\mathcal{H}_t}(m) \leq \left(\frac{em}{t}\right)^t \leq \left(\frac{em}{V(\mathcal{H})}\right)^{tV(\mathcal{H})}.$$

Plugging this and $t \leq m^a$ into an appropriate VC theorem and simplifying (Boucheron et al., 2005a, Theorem 5.1 and subsequent discussion), with probability at least $1 - \delta/2$,

$$\begin{aligned}
R(\hat{A}) &\leq \hat{R}(\hat{A}) + 2\sqrt{\frac{\ln(S_{\mathcal{H}_t}(2m)) + \ln(8/\delta)}{m}} + 4\frac{\ln(S_{\mathcal{H}_t}(2m)) + \ln(8/\delta)}{m} \\
&\leq \hat{R}(\hat{A}) + 2\sqrt{\frac{(V(\mathcal{H})+1)\ln(2em) + \ln(8/\delta)}{m^{1-a}}} + 4\frac{(V(\mathcal{H})+1)\ln(2em) + \ln(8/\delta)}{m^{1-a}}.
\end{aligned}$$

□

Appendix 6.F Deferred Material from Section 6.4

6.F.1 Proof of Proposition 6.4.1

Lemma 6.F.1. Let loss $\ell \in L_{lg}$ and any $g \in \partial\ell(0)$ be given.

1. The restriction of ℓ to $[0, g]$, denoted $\ell_{[0,g]}$, is a (decreasing) bijection between $[0, g]$ and $[-\ell(0), 0]$.
2. Let $f(z) := (\ell_{[0,g]})^{-1}(z)$ denote the inverse of $\ell_{[0,g]}$. If ν is a probability measure, and $p \in L^1(\nu)$, and $r := \int -\ell(p)\, d\nu \in [0, -\ell(0)]$, then $\int [p - f(-r)]_+ \, d\nu = f(-r)$, with $f(-r) > 0$ iff $r > 0$.

Proof. Choose any $g \in \partial\ell(0)$; by Lemma 1.B.1, ℓ is 0 at 0, negative along $(0, g)$, and attains its minimum at g. Since ℓ is finite for every $z \in (0, g)$, then $\partial\ell(z)$ exists (Rockafellar, 1970, Theorem 23.4), and every $x \in \partial\ell(z)$ satisfies $z \in -\partial\ell(x)$ (Rockafellar, 1970, Theorem 23.5), and so ℓ is strictly convex along (Hiriart-Urruty and Lemaréchal, 2001, Theorem E.4.1.2), meaning ℓ is injective along $[0, g]$. Since $\ell(0) = 0$, and $-\ell(g) = -\ell(0)$ (by the Fenchel-Young inequality), and since ℓ is lower semi-continuous (Rockafellar, 1970, Theorem 12.2), then ℓ is also surjective from $[0, g]$ to $[-\ell(0), 0]$.

Now let f denote the inverse map (from $[-\ell(0), 0]$ to $[0, g]$), and let probability measure ν, function $p \in L^1(\nu)$, and scalar $r := \int -\ell(p)\, d\nu \in [0, -\ell(0)]$ be given (where the containment provides $f(-r)$ is valid). By Jensen's inequality,

$$r = \int -\ell(p)\, d\nu \geq -\ell\left(\int p\, d\nu\right);$$

since f is a decreasing map, this implies $\int p \geq f(-r)$. Furthermore,

$$f(-r) \int_{[p<f(-r)/2]} p + \int_{[p \geq f(-r)/2]} 1 \frac{f(-r)}{2} + \int [p - f(-r)/2]_+;$$

meaning $\int [p - f(-r)/2]_+ \geq f(-r)/2$ as desired.

Lastly, the statement $f(-r) > 0$ iff $r > 0$ follows from the bijectivity of $\ell_{[0,g]}$. □

Proof of Proposition 6.4.1. The basic duality relation is provided by Lemma 4.1.1. Since the optimal value satisfies $L \in [0; \ell(0)]$ (since μ is a probability measure, $0 \in \mathcal{H}$ is primal feasible, and $\ell \geq 0$), then Lemma 6.F.1 may be applied with parameter p being the dual optimum and parameter r being the corresponding objective value L. □

6.F.2 Proof of Lemma 6.4.5

The first step is to use p to show that if h has low error and norm, then H will have very small margins over some positive measure set.

Lemma 6.F.2. Let convex differentiable $\ell : \mathbb{R} \to \mathbb{R}_+$ with $\ell'(0) > 0$, any class H, and any probability measure μ over $X \times \{-1, +1\}$ with empirical counterpart $\hat{\mu}$ be given. Suppose the following quantities and constants exist.

1. Suppose there exists $p \in L_1(\mu)$ with $p \in [0, \|p\|_1]$ $\hat{\mu}$-a.e., and that there exists $\gamma > 0$ with $\hat{\mu}([p \leq \gamma]) \leq 1/2$.
2. Set
$$c := \frac{16\ell(0)}{\ell'(0)} \max\left(1, \frac{1}{\gamma}\right) \max\{1, \|p\|_1\},$$
and suppose there exist $C > 0$ and $D \geq c\|p\|_1^2/8$ so that every $h \in \mathcal{H}$ with $\|h\| \leq C$ satisfies $\int_{[yh(x) \leq c]} |h(x)| p \, d\hat{\mu} \leq D$.

If $h \in \mathcal{H}$ satisfies $\|h\| \leq C$ and $\hat{\mu}([yh(x) \leq c]) \leq 1/8$, then $\hat{L}(h) \geq 2\hat{L}(h_0) = 2\ell(0)$.

Proof of Lemma 6.F.2. First consider the case that $\hat{\mu}([yh(x) \leq c]) = \hat{\mu}([h \leq c]) \leq 1/8$. By subgradient rules for convex functions, since $\ell \geq 0$ and $\ell'(0) > 0$, and using the definition of c,

$$\begin{aligned}
\int \ell(h) \, d\hat{\mu} &\geq \int_{[yh(x) \leq c]} \ell(h) \, d\hat{\mu} \\
&\geq \int_{[yh(x) \leq c]} (\ell(0) + \ell'(0) h) \, d\hat{\mu} \\
&\geq \frac{\ell'(0) c}{8} \\
&\geq 2\ell(0),
\end{aligned}$$

meaning $\hat{L}(h) \geq 2\ell(0) = 2\hat{L}(0) = 2L(h_0)$.

Now consider the remaining possibility that $\hat{\mu}([yh(x) \leq c]) < 1/8$. Since by assumption $\hat{\mu}([|yh(x)| \leq c]) \leq 1/8$, it follows that $\hat{\mu}([yh(x) \leq c]) \leq 1/4$. In turn, it also holds

that

$$b([y(H\ )_x \quad c] \setminus [p \quad]) \quad = 4:$$

Next, the de nition of D provides that

$$D \quad \int A\, pd\, b$$
$$\int_{[y(H\ )_x \quad 0]} y(H\ )_x p(x;y) db(x;y) + \int_{[y(H\ )_x > 0]} y(H\ )_x p(x;y) db(x;y);$$

meaning

$$\int_{[y(H\ )_x \quad 0]} y(H\ )_x p(x;y) db(x;y) \quad \int_{[y(H\ )_x > 0]} y(H\ )_x p(x;y) db(x;y) \quad D:$$

As such, since$p \in [0; kpk_1\ ]$ over the sample,

$$k\ pk_1 \int_{[y(H\ )_x \quad 0]} y(H\ )_x db(x;y) \quad \int_{[y(H\ )_x \quad 0]} y(H\ )_x p(x;y) db(x;y)$$
$$\int_{[y(H\ )_x > 0]} y(H\ )_x p(x;y) db(x;y) \quad D$$
$$\int_{[y(H\ )_x \quad c] \setminus [p \quad]} y(H\ )_x p(x;y) db(x;y) \quad D$$
$$\frac{c^{\ 2}}{4} \quad D:$$

Turning back to $\hat{b}$ and proceeding similarly to the earlier case,

$$\int \grave{}(A\ ) db \quad \int_{[y(H\ )_x \quad 0]} \grave{}(A\ ) db$$
$$\int_{[y(H\ )_x \quad 0]} (\grave{}(0) + \grave{}'(0)(A\ ))\, db$$
$$\frac{\grave{}'(0)}{kpk_1} \quad \frac{c^{\ 2}}{4} \quad D$$
$$2\grave{}(0);$$

which again yields $L(A\ ) \quad 2\grave{}(0) = 2L(0) = 2L(A_{\ 0})$. □

Next, these small margins in turn cause the line search to not look too far, meaning the next iterate will also have some small margins. Note that$H$ is assumed binary; this is in order to

changes in H to changes in ℓ.

Lemma 6.F.3. Let convex $\ell : R \to R_+$, binary H, and probability measure ℓ with empirical counterpart $\hat{\ell}$ be given. Let positive reals C, c, ℓ be given so that $\ell 2$ with $\|\ell - \ell_1\| \le C + 2c$ and $\hat{b}([|H \ell - \ell| \ge c]) \le 1 - \ell/8$ implies $L_{\hat{b}}(A \ell) \le \ell(0)$. Then for any $\ell 2$ with $\|\ell\| \le \|\ell_1\| \le C$ and $L(A \ell) \le \ell(0)$, the set of line search candidates

$$S := \left\{ \ell' 2 : L_{\hat{b}}(A \ell') < \ell(0), \exists \ell 2 R, h 2 H \ \ell' := \ell + \ell e_h \right\}$$

satisfies $\hat{b}([|H \ell' - \ell| \ge c]) < 1 - \ell/8$ for every $\ell' 2 S$.

Proof. Let ℓ with $\|\ell\| \le \|\ell_1\| \le C$ be given, and consider any ℓ' of the form $\ell' := \ell + \ell' e_h$ for some $\ell' 2 R$ and $h 2 H$. The desired statement will be shown by contrapositive; namely, $\hat{b}([|H \ell' - \ell| \ge c]) \le 1 - \ell/8$ implies $\ell' 62 S$.

For any example (x, y), since H has binary predictors, the map $\ell \mapsto 1[|y(H(\ell + \ell e_h))_x| \ge c]$ is constant for $\ell > 2c$ and for $\ell < 2c$; consequently, since $\hat{b}$ is a discrete measure over a finite set, the map $\ell \mapsto \hat{b}([|y((H(\ell + \ell e_h))_x| \ge c])$ is also constant for $\ell > 2c$ and $\ell < 2c$. As such, the existence of ℓ' as above implies the existence of $\ell'' := \ell + \ell'' e_h$ where $h 2 H$ is as before, ℓ'' and ℓ' have the same sign, and $|\ell''| \le \min\{2c, |\ell'|\}$; in other words, ℓ'' is along the path from ℓ to ℓ', but moreover satisfies $\|\ell'' - \ell\| \le 2c$. But this means $\|\ell''\| \le \|\ell\| + \|\ell'' - \ell\| \le \|\ell_1\| + 2c \le C + 2c$, whereby the stated assumptions combined with $\hat{b}([|H \ell' - \ell| \ge c]) \le 1 - \ell/8$ provide $L_{\hat{b}}(A \ell'') \le \ell(0)$, thus $\ell'' 62 S$. Furthermore, since ℓ'' is along the path from ℓ to ℓ', and $L_{\hat{b}}(A \ell) \le \ell(0)$ and $L_{\hat{b}}(A \ell'') \le \ell(0) \le L_{\hat{b}}(A \ell)$, it follows by convexity that $L_{\hat{b}}(A \ell') \le \ell(0)$, and thus $\ell' 62 S$ as well. □

These small margin controls directly give a bound on step sizes.

Lemma 6.F.4. (See also Bartlett and Traskin (2007, eq. (28)).) Let $\ell 2 L_{lg} \cap L_{2d}$ be given with Lipschitz gradient parameter B_2, binary class H, time horizon t, and empirical probability measure $\hat{b}$ corresponding to a sample of size m be given. Let positive real $B_1 > 0$ be given so that for any $\ell 2$ with

$$\|\ell\| \le \|\ell_1\| + \sqrt{\frac{t \max\{\ell, 2B_2 = B_1 g(L_{\hat{b}}(A \ell_0) - L_{\hat{b}}(A \ell_t))}{\ell^2 B_1}},$$

and any line search candidate $\ell' := \ell + \ell e_h$ for $h 2 H$, $\ell 2 R$, and satisfying $L_{\hat{b}}(A \ell') \le L_{\hat{b}}(A \ell)$, then $B_1 \le \frac{1}{m} \sum_{i=1}^{m} \ell''((A \ell')_i)$. The following properties hold.

1. For every integer $0 \le i < t$, an optimal step η_i exists, and every step sizes choice satisfies

$$\eta_i^2 \le \min\left(\frac{9\|A^\top \nabla \hat{L}(A_{i-1})\|_1^2}{4\beta B_1^2}\,;\ \frac{\max\{5, 2B_2/B_1\}(\hat{L}(A_{i-1}) - \hat{L}(A_i))}{\beta B_1} \right).$$

2. For every integer $0 \le i < t$ and any sequence of step size choices,

$$\|A_i\|_1 \le \sqrt{\sum_{j=1}^{i} \eta_j^2}\ \sqrt{\frac{\max\{5, 2B_2/B_1\}(\hat{L}(A_0) - \hat{L}(A_i))}{\beta B_1}} \le \sqrt{\frac{\ell(0)\max\{5, 2B_2/B_1\}}{\beta B_1}}.$$

Proof. This proof establishes both properties simultaneously by induction on i. In the base case $\eta_i = \eta_0 = 0$ and there is nothing to show, thus suppose $i \ge 1$.

Define the interval $I_i := \{\eta \ge 0 : \hat{L}(A(\eta_{i-1} + \eta v_i)) \le \hat{L}(A_{i-1})\}$. Combining the inductive hypothesis (controlling $\|A_{i-1}\|_1$) with the assumptions on line search candidates means the second-order lower bound B_1 is active along I_i. Now consider a Taylor expansion of $\hat{L}$, but in the direction reverse to Lemma 1.C.3, and using the fact that H is binary; then for any $\eta \in I_i$ and some $z \in [\eta_{i-1}, \eta_{i-1} + \eta v_i]$,

$$\begin{aligned}\hat{L}(A(\eta_{i-1} + \eta v_i)) &= \hat{L}(A_{i-1}) + \eta \left\langle A^\top \nabla \hat{L}(A_{i-1}), v_i \right\rangle + \frac{\eta^2}{2m} \sum_i \ell''((Az)_i)(Av_i)^2 \\ &\le \hat{L}(A_{i-1}) - \eta \|A^\top \nabla \hat{L}(A_{i-1})\|_1 + \frac{\eta^2}{2} B_1.\end{aligned}$$

This last expression defines a univariate quadratic which lies below $\hat{L}(A(\eta_{i-1} + \eta v_i))$ along I_i (outside of I_i, the constraints granting the lower bound B_1 may be violated). Consequently, I_i is bounded, and the optimal step η_i must exist, and moreover satisfies

$$\eta_i \le \eta := \frac{\|A^\top \nabla \hat{L}(A_{i-1})\|_1}{B_1}; \tag{6.F.5}$$

where η is the minimizer to the above quadratic. Plugging η back into the quadratic, for any $\eta \in I_i$,

$$\hat{L}(A(\eta_{i-1} + \eta v_i)) \le \hat{L}(A_{i-1}) - \frac{\|A^\top \nabla \hat{L}(A_{i-1})\|_1^2}{2B_1}. \tag{6.F.6}$$

Now consider the first two step size options in Figure 1.3; combining eq. (6.F.6) with

Lemma 1.C.3 and the Lipschitz gradient guarantee in Lemma 6.B.2,

$$\frac{2}{i} \quad \frac{2}{i} \quad \frac{kA^{>} \hat{L}b(A_{i\ 1})k_1^2}{B_1^2} \quad \frac{2B_2(\hat{L}b(A_{i\ 1}) \quad \hat{L}b(A_i))}{^2B_1^2}$$

as desired.

For option 3 (the Wolfe search), combining eq. (1.C.5) with eq. (6.F.6) grants

$$\frac{2}{i} \quad \frac{9(\hat{L}b(A_{i\ 1}) \quad \hat{L}b(A_i))^2}{^2kA^{>} r\ \hat{L}b(A_{i\ 1})k_1^2} \quad \frac{9(\hat{L}b(A_{i\ 1}) \quad \hat{L}b(A_i))}{2\ ^2B_1} \quad \frac{9kA^{>} r\ \hat{L}b(A_{i\ 1})k_1^2}{4\ ^2B_1^2};$$

which establishes the first inductive property for all step sizes.

The second statement is just Cauchy-Schwarz combined with the bound on $\frac{2}{i}$:

$$\begin{aligned} k\ _i k_1 &\quad \frac{\sum_{j=1}^{i} {}_j kv_j k_1}{\sqrt{i}\sqrt{\sum_{j=1}^{i} {}_j^2}} \\ &\quad \sqrt{i}\sqrt{\frac{\max\{5; 2B_2/B_1\} \sum_{j=1}^{i} (\hat{L}b(A_{j\ 1}) \quad \hat{L}b(A_j))}{^2B_1}} \\ &= \sqrt{i}\sqrt{\frac{\max\{5; 2B_2/B_1\}(\hat{L}b(A_0) \quad \hat{L}b(A_i))}{^2B_1}}; \end{aligned}$$

and nonnegativity of $\grave{}$ and the fact that all steps perform descent grants $\hat{L}b(A_0) \quad \hat{L}b(A_i)$ $\grave{}(0)$. □

After some algebra, the upper bound also grants a lower bound; due to this indirection, it should be possible to improve this bound. Note that the beginning of this derivation, when initially lower bounding $_i$, uses derivations similar to those used by Zhang and Yu (2005) and Bartlett and Traskin (2007).

Lemma 6.F.7. Let $\grave{} \ 2\ L_{lg} \setminus L_{2d}$ be given with Lipschitz gradient parameter B_2, binary class H, time horizon t, and empirical probability measure b corresponding to a sample of size n be given. Suppose there exists $c_2 > 0$ with $k\ _i k_1 \quad c_2\sqrt{i}$ for all $0 \quad i \quad t$. Additionally, let $C_2 > 0$ and

2 be given with $k \ k_1 \quad C_2$ and $:= \min_{i \ 2 \ [t \ 1]} \hat{L}(A_{\ i}) \quad \hat{L}(A \ ) \quad 0$. Then

$$_i \quad \frac{(\hat{L}(A_{\ i \ 1}) \quad \hat{L}(A \ ))}{2B_2(k \ k_1 + c_2 \sqrt{i \quad 1})}$$

and

$$\sum_{i=1}^{t} \ _i \quad \frac{}{2B_2c_2} \quad 2\sqrt{t \quad 1} + \frac{c_2}{C_2} + \frac{2C_2}{c_2} \ln \quad \frac{C_2}{C_2 + c_2\sqrt{t \quad 1}} \quad ;$$

or more simply $\sum_{i=1}^{t} \ _i \quad (\ \sqrt{t \quad 1})/(4B_2c_2)$ if $C_2 \quad c_2\sqrt{t \quad 1}$.

Proof. For any $1 \quad i \quad t$, note by Lemma 1.A.5 that

$$\begin{aligned} \|A^{>} r \ \hat{L}(A_{\ i \ 1})\|_{k_1} &= \sup_{k \ k_1 \ 1} \ \left\langle A^{>} r \ \hat{L}(A_{\ i \ 1}), \right. \\ &\quad \left. \frac{1}{k_{\ i \ 1} \quad k_1} \ r \ \hat{L}(A_{\ i \ 1}); A(\ _{i \ 1} \quad) \right\rangle \\ &\quad \frac{1}{k_{\ i \ 1} \quad k_1} (\hat{L}(A_{\ i \ 1}) \quad \hat{L}(A \ )): \end{aligned}$$

Thus, by the lower bounds in Lemmas 1.C.3 and 1.C.7 for every step size (combined with the Lipschitz gradient guarantee in Lemma 6.B.2),

$$_i \quad \frac{\|A^{>} r \ \hat{L}(A_{\ i \ 1})\|_{k_1}}{2B_2} \quad \frac{(\hat{L}(A_{\ i \ 1}) \quad \hat{L}(A \ ))}{2B_2k_{\ i \ 1} \quad k_1} \quad \frac{}{2B_2k_{\ i \ 1} \quad k_1}:$$

Combining this with the provided upper bound on $k \ _i k_1$,

$$_i \quad \frac{}{2B_2(k \ k_1 + k_{\ i \ 1}k_1)} \quad \frac{}{2B_2(k \ k_1 + c_2\sqrt{i \quad 1})}$$

As a consequence of this, and recalling the simplification $k \ k_1 \quad C_2$,

$$\begin{aligned} \sum_{i=1}^{t} \ _i \quad & \frac{}{2B_2c_2} \quad \frac{c_2}{C_2} + \sum_{i=1}^{t \ 1} \frac{c_2}{C_2 + c_2\sqrt{i}} \\ & \frac{}{2B_2c_2} \quad \frac{c_2}{C_2} + \int_0^{t \ 1} \frac{c_2 dx}{C_2 + c_2\sqrt{x}} \\ = & \frac{}{2B_2c_2} \quad \frac{c_2}{C_2} + 2\sqrt{x} \quad \frac{2C_2 \ln(\sqrt{x} + C_2/c_2)}{c_2} \Bigg|_0^{t \ 1} \\ = & \frac{}{2B_2c_2} \quad \frac{c_2}{C_2} + 2\sqrt{t \quad 1} + \frac{2C_2}{c_2} \ln \quad \frac{C_2}{C_2 + c_2\sqrt{t \quad 1}} \quad : \end{aligned}$$

When $C_2 \quad c_2 \sqrt{t \quad 1}$, it su ces to instantiate the above bound with $C_2^0 := c_2 \sqrt{t \quad 1}$ (whereby it still holds that $k \ k_1 \quad C_2 \quad C_2^0$), and then rearrange, noting $1 \quad \ln(2) \quad 1=4$ and deleting the nonnegative standalone term $c_2=C_2$. □

By combining the above chain of results, the proof of Lemma 6.4.5 follows.

Proof of Lemma 6.4.5. Recalling the structure from Proposition 6.4.1, let dual optimum p and real > 0 be given so that $([p \quad]) \quad$. By Hoe ding's inequality and the rst lower bound on m, with probability at least $1 \quad =4$

$$b([p \quad]) \quad ([p \quad]) \quad \sqrt{\frac{1}{2m} \ln \frac{4}{}} \quad \frac{}{2}:$$

Henceforth disregard the corresponding failure event.

Next de ne

$$D := \frac{2(R_t + 2c)kpk_1}{m^{1=2}} \quad 2^p \overline{2V(H)\ln(m+1)} + \sqrt{2\ln(4=)} \quad :$$

The second condition on m grants $D \quad c^{\ 2}=8$, whereby Lemma 6.4.4 grants, with probability at least $1 \quad =4$, that every 2 with $k \ k_1 \quad R_t + 4c$ satis es

$$\int (A \ )pdb \quad D \quad \frac{c^{\sim 2}}{8}:$$

Discard the corresponding failure event as well; unioning this with the earlier failure event, the remaining steps hold with probability at least $1 \quad =2$.

It follows from Lemma 6.F.2 that every such 2 with $k \ k_1 \quad R_t + 2c$ either satis es $b([jH \ j \quad c]) < 1 \quad \sim=8$, or else $\underline{b}(A \ ) \quad 2`(0)$. This in turn means the preconditions to Lemma 6.F.3 are met (with $C := R_t$), and in particular, for any 2 with $k \ k_1 \quad R_t$ and $\underline{b}(A \ ) \quad \underline{b}(A \ _0)$, the set of line search candidates

$$S := \left\{ {}^0 2 \ : \ \underline{b}(A \ ^0) < 2`(0); 9 \ 2 \ R; h \ 2 \ H \quad {}^0 := \quad + \ e_h \right\}$$

satis es $b([jH^{0}j \quad c]) < 1 \quad =8$ for every $^{0}2 S$. But then, for every $^{0}2 S$ (and note $2 S$),

$$\frac{1}{m}\sum_{i2[m]} `^{00}((A)_{x_i;y_i}) \quad \frac{1}{m}\sum_{i2[m]:\, j(A)_{x_i;y_i}j \ c} `^{00}((A)_{x_i;y_i}) \quad b([jH \ j \ c]) \inf_{z2[\ c;+c]} `^{00}(z) \quad B_1:$$

This establishes the rst desired statement.

For the second statement (upper bounds on $_i$ and $k \ _ik_1$), note that the above properties satisfy the preconditions to Lemma 6.F.4, whereby the desired upper bounds follow.

Similarly, for the third statement (lower bounds on $_i$ and $k \ _ik_1$), the preconditions for Lemma 6.F.7 are now met. □

6.F.3 Optimization Guarantees

As stated previously, the following proof is a reworking of a proof due to Zhang and Yu (2005), albeit with the present decoupling of line search and coordinate selection.

Proof of Lemma 6.4.6. Let $1 \ i \ t$ be arbitrary. The rst step of this proof is to develop two lower bounds on $kA^{>}r \ L\!b(A_{\ i \ 1})k_1$. First, just as in the proof of Lemma 6.F.7,

$$\begin{aligned} kA^{>}r \ L\!b(A_{\ i \ 1})k_1 &= \sup_{k \ k_1 \ 1} \left\{ \left\langle A^{>}r \ L\!b(A_{\ i \ 1}); \ \right\rangle \right. \\ &\quad \left. \frac{1}{k_{\ i \ 1} \quad k_1} \left\langle r \ L\!b(A_{\ i \ 1}); A(\ _{i \ 1} \quad) \right\rangle \right\} \\ &\quad \frac{1}{k_{\ i \ 1} \quad k_1}(L\!b(A_{\ i \ 1}) \quad L\!b(A)): \end{aligned}$$

The second lower bound is provided by assumption, and thus, by the guarantee on any line search as in Lemmas 1.C.3 and 1.C.7 (together with the Lipschitz gradient property in Lemma 6.B.2),

$$\begin{aligned} L\!b(A_{\ i}) \quad L\!b(A) \quad L\!b(A_{\ i \ 1}) \quad L\!b(A) \quad & \frac{^{2}kA^{>}r \ L\!b(A_{\ i \ 1})k_1^2}{6B_2} \\ L\!b(A_{\ i \ 1}) \quad L\!b(A) \quad & \frac{c_3 \ ^{2}{}_i(L\!b(A_{\ i \ 1}) \quad L\!b(A))}{6B_2k_{\ \ i \ 1}k_1} \\ = \ L\!b(A_{\ i \ 1}) \quad L\!b(A) \quad 1 \quad & \frac{c_3 \ ^{2}{}_i}{6B_2k_{\ \ i \ 1}k_1} \\ L\!b(A_{\ i \ 1}) \quad L\!b(A) \ \exp \quad & \frac{c_4 \ _i}{k_{\ \ i \ 1}k_1} : \end{aligned}$$

where the last step took $c_4 := c_3^{\ 2}/(6B_2)$ for convenience. Iterating this bound,

$$\underline{b}(A_t) \quad \underline{b}(A) \qquad \underline{b}(A_0) \quad \underline{b}(A) \quad \exp\left(c_4 \sum_{i=1}^{t} \frac{i}{k_{\ i\ 1} k_1} \right) :$$

Focusing on the summation, define $S_i := \sum_{j\ i\ i}$ with $S_0 := 0$, whereby $k_{\ i} k_1 \quad S_i$. Using this (see also the similar derivation by Zhang and Yu (2005, Proof of Lemma 4.2)),

$$\begin{aligned}
\sum_{i=1}^{t} \frac{i}{k\ k_1 + k_{\ i\ 1} k_1} &\quad \sum_{i=1}^{t} \frac{i}{k\ k_1 + S_{i\ 1}} \\
&= \sum_{i=1}^{t} \frac{k\ k_1 + S_i}{k\ k_1 + S_{i\ 1}} \quad \frac{k\ k_1 + S_{i\ 1}}{k\ k_1 + S_{i\ 1}} \\
&\sum_{i=1}^{t} \ln \frac{k\ k_1 + S_i}{k\ k_1 + S_{i\ 1}} \quad = \ln \frac{k\ k_1 + S_i}{k\ k_1} :
\end{aligned}$$

Plugging this into the preceding display and collecting terms, the result follows. □

Note that the substitution $k_{\ i} k_1 \quad S_i$ at the end of the proof of Lemma 6.4.6 works around the fact that $\sum_{i\ i}$ could be much larger than $k_{\ i} k_1$; this issue is frequently avoided in the literature by assuming that H is closed under negation, whereby $v_i = e_{h_i}$ in each round (i.e., rather than $v_i \ 2 \ f \quad e_{h_i} g$).

6.F.4 Statistical Guarantees

Proof of Theorem 6.4.2. Let > 0 be arbitrary, and choose 2 with $k\ k_1 \quad R_{t\ 1}$ so that

$$L(A) \qquad + \inf_{k\ k_1 \ R_{t\ 1}} L(A) :$$

By McDiarmid's inequality and the fact that $\sup_{x;y} j(A)_{x;y} j \quad R_{t\ 1}$, with probability at least $1 \quad =6$,

$$\underline{b}(A) \quad R_{t\ 1} \sqrt{\frac{2}{m} \ln \frac{6}{}} + L(A) \qquad + R_{t\ 1} \sqrt{\frac{2}{m} \ln \frac{6}{}} + \inf_{k\ k_1 \ R_{t\ 1}} L(A) :$$

Now let $b \ 2 \ (0; 1=2)$ be arbitrary, and consider two cases for the difference $:= \min_{i \ 2 [t \ 1]} \underline{b}(A_{\ i}) \quad \underline{b}(A)$.

If $(t \quad 1)^{\ b}$, then

$$\hat{L}(A_{\ t}) \quad \hat{L}(A_{\ t\ 1}) \quad (t \quad 1)^{\ b} + \hat{L}(A\ ):$$

Otherwise $> (t \quad 1)^{\ b}$. Thus Lemma 6.4.5 grants, with probability at least $1 \quad =2$,

$$_i^2 \quad \frac{9kA^{>}r\ \hat{L}(A_{\ i\ 1})k_1^2}{4\ ^2B_1^2} \qquad \text{and} \qquad \sum_{i=1}^{t} \quad _i \quad \frac{p\overline{t\ \ 1}}{4B_2R_1};$$

which can then be plugged into Lemma 6.4.6 (with $c_3 := 2\ B_{\ 1}=3$), together with the lower bound on , to yield

$$\hat{L}(A_{\ t}) \quad \hat{L}(A\ ) \quad \hat{L}(A_{\ 0}) \quad \hat{L}(A\ ) \quad \frac{k\ k_1}{k\ k_1 + \sum_{i\ t\ i}}^{!\ c_3\ ^2=(6B_2)}$$

$$\hat{L}(A_{\ 0}) \quad \hat{L}(A\ ) \quad \frac{k\ k_1}{k\ k_1 + \ (t \quad 1)^{1=2\ \ b}=(4B_2R_1)}^{\ ^3B_1=(9B_2)}:$$

Summing these two bounds gives a relation which holds in general; consequently, with probability at least $1 \quad =2$ (due to the invocation of Lemma 6.4.5),

$$\hat{L}(A_{\ t}) \quad \hat{L}(A\ ) \quad (t \quad 1)^{\ b} + \grave{}(0) \quad \frac{k\ k_1}{k\ k_1 + \ (t \quad 1)^{1=2\ \ b}=(4B_2R_1)}^{\ ^3B_1=(9B_2)}:$$

The rst desired claim follows (with probability $1 \quad 5=6$) by recalling the earlier application of McDiarmid's inequality to , combined with $\# 0$, the choice $b := 1=4$ (which is not optimal, but neither is the exponent $B_1\ ^3=(9B_2)$), and standard Rademacher bounds for voting classi ers applied to Lipschitz losses (Boucheron et al., 2005a, Theorem 4.1, eq. (8), and their proofs, which control for L), which makes use of the bound $k\ _tk_1 \quad R_t$ (granted by the earlier instantiation of Lemma 6.4.5), and simplifying via $t \quad 1 \quad t$,

$$L(A_{\ t}) \quad \inf_{k\ k_1\ \ R_{t\ \ 1}} \hat{L}(A\ ) + t^{\ 1=4} + R_{t\ \ 1}\ \sqrt{\frac{2}{m}\ln\ \frac{6}{\delta}}$$

$$+ \frac{2\ _2R_t}{m^{1=2}}\ \ 2^{\sqrt{2V(H)\ln(m+1)}} + \grave{}(R_t)\sqrt{2\ln(6=\ )}$$

$$+ \grave{}(0)\ \ \frac{k\ k_1}{k\ k_1 + \ t^{\ 1=4}=(4B_2R_1)}^{\ ^3B_1=(9B_2)}:$$

Plugging in $t = m^{\alpha}$ gives the first result.

For the second guarantee, since $H \in \mathcal{F}_{ds}(\mathcal{X})$, then Lemma 6.A.1 grants

$$\inf\{L(A) : \quad \geq \quad g = \inf \int \ell(\quad yf(x))\, d\,(x,y) : f \in \mathcal{F}_b \quad ;$$

where $\mathcal{F}_b$ is the family of Borel measurable functions from $\mathcal{X}$ to $\mathbb{R}$. From here, since ℓ is classification calibrated (Bartlett et al., 2006, Theorem 2, noting that the present thesis instead takes losses to be nondecreasing rather than nonincreasing), there exists a function $\quad : \mathbb{R} \to \mathbb{R}$ satisfying

$$R(A_t) \quad R^{*} \quad \quad L(A_t) \quad \inf_{f \in \mathcal{F}_b} \int \ell(\quad yf(x))\, d\,(x,y) = L(A_t) \quad \inf_{2} L(A)$$

(where the last step used the previous display), and moreover $(z) \to 0$ as $z \to 0$ (Bartlett et al., 2006, Theorem 1). The specialization for $\ell_{\log}$ is due to Zhang (2004, Subsection 3.5 and Corollary 3.1).

The final guarantee follows by applying a version of the VC theorem to the predictors, and follows the exact strategy as in Theorem 6.3.2 (but using failure probability $=6$), and making use of the equality

$$R(\hat{H}) \quad R^{*}(H_t) = (R(\hat{H}) \quad \hat{R}(\hat{H})) + (\hat{R}(\hat{H}) \quad \hat{R}(H_t)) + (\hat{R}(H_t) \quad R^{*}(H_t));$$

and the fact that $\hat{R}(\hat{H}) \quad \hat{R}(H_t)$. □

Appendix 6.G Proof of Consistency

Proof of Theorem 6.2.2. This proof is a standard application of the Borel-Cantelli Lemma; for an exposition on such applications, please see the proof of consistency of AdaBoost due to Schapire and Freund (2012, Proof of Corollary 12.3). In particular, let any > 0 be given, and let $E_{m,}$ be the event that the output $\hat{}_m$, trained on m examples, has classification risk $R(\hat{H}_m)$ exceeding the Bayes risk $R_?$ by more than ; to prove consistency, it suffices (thanks to Borel-Cantelli) to show that $\sum_m \Pr(E_{m,}) < 1$.

There are two cases to consider: either $L = 0$, or $L > 0$. In the case that $L = 0$,

instantiate the nite sample guarantee in Theorem 6.3.2 for each $m \quad 1$ with $= 2$ and $:= 1=m^2$; there is a real $M < 1$ where $m > M$ provides the preconditions on the bound are met and the bound is at most , and thus

$$\sum_{m=1}^{1} \Pr(E_{m;}) \quad \sum_{m=1}^{M} 1 + \sum_{m=M+1}^{1} \frac{1}{m^2} \quad M + \frac{2}{6} < 1 :$$

When $L > 0$, once again instantiate a relevant nite sample guarantee, this time from Theorem 6.4.2, with $:= 1=m^2$. It will be necessary to use all three guarantees; rst, let M_1 be su ciently large so that the third guarantee provides $R(H\hat{}_m) \quad R(H^{0}_{m}) + =2$ with failure probability m^{2} for all $m > M_1$, where $^{0}_{m}$ is the last iterate considered by the algorithm when run on m examples (thus $^{0}_{m}$ is basically $_{m^a}$, modulo rounding issues). Next, by the second guarantee, there exists ^{0} small enough so that

$$^{0} \quad (R(A^{0}_{m}) \quad R_{?}) \quad =) \quad R(A^{0}_{m}) \quad R_{?}:$$

As such, now consider the rst guarantee, where the goal will be to establish that $L(A^{0}_{m}) \quad L^{0}$ for all large m. But note rstly that the quantity $R_{t\ 1} ! 1$ as $m ! 1$, which combined with

$$\inf\{L(A) : \ 2 \ g = \inf_{C>0} \inf\{L(A) : \ 2 \ ; k\ k_1 \ C g;$$

grants that, if $m > M_2$ (for some M_2), then $\inf_{k\ k\ R_t} L(A) \quad L + {}^{0}=2$. Finally, the rest of the terms in the rst guarantee are at most $^{0}=2$ for $m > M_3$ (for some M_3) with the same m^{2} failure probability. As such, similarly to before, $\sum_m \Pr(E_{m;}) \quad \max\{M_1; M_2; M_3\} + {}^{2}=6 < 1$. □

Appendix 6.H Bibliographic Notes

The consistency of AdaBoost was rst analyzed under various regularization strategies. Most notably, the work of Blanchard et al. (2003) and Lugosi and Vayatis (2004) studied the solutions of penalized estimators; the former work in particular achieving excellent nite sample guarantees, with convex risk decaying roughly as $O(m^{1=2})$ (where m is the sample size), with improvements under various noise conditions. This work, however, did not demonstrate tractable algorithms to produce these estimators, which was a goal of the work by Zhang and Yu (2005);

namely, there it is shown that merely constraining the step size taken by an AdaBoost-style scheme (with a variety of losses) su ces to achieve a convex risk rate of roughly $O(m^{-1/4})$ (in the case of the logistic loss), which includes the e ect of approximate solutions produced by the algorithm. As will be discussed later, the present chapter, in the nonseparable case, ts well with the development by Zhang and Yu (2005).

Two works give a consistency analysis of AdaBoost without any algorithmic modi cations, under the condition that the algorithm is stopped after m^a iterations (with arbitrary $a \in (0,1)$). The rst such analysis, due to Bartlett and Traskin (2007), was focused on establishing consistency, and established the convex risk decays roughly as $O(1/\sqrt{\ln(m)})$; the analysis depends on a curvature lower bound, which follows from a lower bound on the convex risk since the exponential loss is equal to its derivative. This derivative structure is of course not present with the logistic loss, and the analysis in this chapter must nd another way. A streamlined consistency analysis of AdaBoost appears in the textbook of Schapire and Freund (2012, Theorem 12.2), with a rate of roughly $O(m^{-1/9})$ (by choosing $a := 5/9$); the analysis is short and clean, but it is not clear how to decouple the exponential loss.

In the separable case, the analysis in this chapter relies upon ideas from weak learnability, just as with the original analysis of AdaBoost (under margin assumptions) (Freund and Schapire, 1997). The relaxed notion of margin here is very close to the quantity AvgMin_k as developed by Shalev-Shwartz and Singer (2008, Section 4.1); the main contrasting point is that this thesis chapter is concerned with statistical properties, and in particular how these relaxed margin properties behave under sampling. The optimization analysis in the separable case here shares ideas both with the original AdaBoost analysis (Freund and Schapire, 1997), but also with the literature on hard cores (Impagliazzo, 1995, Barak et al., 2009); one distinction is that the latter methods take the target weak learning rate as input, whereas here (and in general with adaptive boosting), it must be found by the algorithm. Interestingly, the loss function implicit in the boosting algorithm due to Impagliazzo (1995, Proof of Lemma 1), $\grave{}_{\mathrm{russ}}$ as discussed in Section 1.3, achieves superior constants to the logistic loss in Theorem 6.3.2; nearly the same loss was presented and praised by Zhang (2004, see the de nition at the end of Section 4.6).

As stated previously, the nonseparable case ts well with the scheme laid down by Zhang and Yu (2005), where the algorithm is modi ed to constrain step sizes. Indeed, the analysis here rst establishes that the iterates are well-behaved with exactly the sorts of norm bounds needed

by the analysis of Zhang and Yu (2005) (compare for instance the summability conditions (Zhang and Yu, 2005, Equation (4)) with Lemma 6.4.5). In order to produce these results, the present work uses a dual optimum as a witness to the difficulty of the convex risk problem over the source distribution; this technique follows structural properties extrapolating from the finite setting in Chapter 2. The corresponding statistical analysis in Chapter 3 appears statistically unstable, and the subsequent analysis here follows a similar path to the one by Zhang and Yu (2005), with additional help from Bartlett and Traskin (2007). One interesting distinction between the present work and those by Bartlett and Traskin (2007) and Zhang and Yu (2005) is that the latter two require a more strenuous algorithm: the weak learner and step size selection must be performed simultaneously. Decoupling these does not appear to impact the rates, however, this distinction prevents those results from being directly invoked here, meaning they must instead be reworked.

Lastly, note that the translation between convex and classification risks follows standard results on classification calibration as first developed by Zhang (2004), and later extended by Bartlett et al. (2006).

Acknowledgements

This chapter is based on work by the dissertation author which appeared in the Conference on Learning Theory, 2013.

Bibliography

Boaz Barak, Moritz Hardt, and Satyen Kale. The uniform hardcore lemma via approximate bregman projections. In SODA, pages 1193{1200, 2009.

Peter L. Bartlett and Mikhail Traskin. AdaBoost is consistent. Journal of Machine Learning Research, 8:2347{2368, 2007.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuli e. Convexity, classi cation, and risk bounds. Journal of the American Statistical Association, 101(473):138{156, 2006.

Adi Ben-Israel. Motzkin's transposition theorem, and the related theorems of Farkas, Gordan and Stiemke. In M. Hazewinkel, editor, Encyclopaedia of Mathematics, Supplement III. 2002.

Dimitri P. Bertsekas. Nonlinear Programming. Athena Scienti c, 2 edition, 1999.

Peter J. Bickel, Yaacov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. Journal of Machine Learning Research, 7:705{732, 2006.

Gilles Blanchard, Gabor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classi ers. Journal of Machine Learning Research, 4:861{894, 2003.

Jonathan Borwein and Adrian Lewis. Convex Analysis and Nonlinear Optimization. Springer Publishing Company, Incorporated, 2000.

Jonathan M. Borwein and Qiji J. Zhu. Techniques of Variational Analysis. Springer, 1 edition, 2005.

Stephane Boucheron, Olivier Bousquet, and Gabor Lugosi. Theory of classi cation: a survey of recent advances. ESAIM: Probability and Statistics, 9:323{375, 2005a.

Stephane Boucheron, Olivier Bousquet, and Gabor Lugosi. Theory of classi cation: A survey of some recent advances. ESAIM: Probability and Statistics, 9:323{375, 2005b.

Leo Breiman. Prediction games and arcing algorithms. Neural Computation, 11:1493{1517, October 1999.

Leo Breiman. Some in nity theory for predictor ensembles, 2000. Berkeley statistics technical report 577.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. pages 161{168, 2006.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. Machine Learning, 48(1-3):253{285, 2002.

George B. Dantzig and Mukund N. Thapa. Linear Programming 2: Theory and Extensions Springer, 2003.

Sanjoy Dasgupta and Phil Long. Boosting with diverse base classi ers. In COLT, 2003.

L. Devroye, L. Gyor , and G. Lugosi. A probabilistic theory of pattern recognition. Springer, 1996.

Nigel Du y and David Helmbold. Potential boosters? In NIPS, pages 258{264. MIT Press, 2000.

Rick Durrett. Probability: Theory and Examples Cambridge University Press, 4 edition, 2010.

Gerald B. Folland. Real analysis: modern techniques and their applications Wiley Interscience, 2 edition, 1999.

Yoav Freund. Boosting a weak learning algorithm by majority. Information and Computation, 121(2):256{285, 1995.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci., 55(1):119{139, 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28(2):337{407, 2000.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189{1232, 2000.

Oded Goldreich and Leonid Levin. A hard-core predicate for all one-way functions. STOC, pages 25{32, 1989.

Christian Gourieroux and Alain Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. Journal of Econometrics, 17(1):83{97, 1981.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In FOCS, 2006.

Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. Fundamentals of Convex Analysis Springer Publishing Company, Incorporated, 2001.

Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In FOCS, pages 538{545, 1995.

Michael Kearns and Leslie Valiant. Cryptographic limitations on learning nite automata and boolean formulae. STOC, pages 433{444, 1989.

Michael Kearns and Umesh Vazirani. An introduction to computational learning theory. MIT Press, 1994.

Jyrki Kivinen and Manfred K. Warmuth. Boosting as entropy projection. In COLT, pages

134{144, 1999.

Hidetoshi Komiya. Elementary proof for sion's minimax theorem. Kodai Mathematical Journal, 11(1):5{7, 1988.

John La erty. Additive models, boosting, and inference for generalized divergences. In In Proc. 12th Annu. Conf. on Comput. Learning Theory, pages 125{133. ACM Press, 1999.

Guy Lebanon. Consistency of the maximum likelihood estimator, 2008. URL http://www.cc. gatech.edu/ lebanon/notes/mleConsistency.pdf.

Phil Long and Rocco Servedio. Algorithms and hardness results for parallel large margin learning. In NIPS, 2011.

Gabor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. Annals of Statistics, 32(1):30{55, 2004.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research, 46(1):157{178, 1993.

Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus R. Frean. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Scholkopf, and D. Schurmans, editors, Advances in Large Margin Classi ers, pages 221{246, Cambridge, MA, 2000. MIT Press.

Indraneel Mukherjee, Cynthia Rudin, and Robert Schapire. The convergence rate of AdaBoost. In COLT, 2011.

Yurii Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Springer, 1 edition, 2003.

Jorge Nocedal and Stephen J. Wright. Numerical optimization. Springer, 2 edition, 2006.

Gunnar Ratsch and Manfred Warmuth. E cient margin maximizing with boosting. Journal of Machine Learning Research, 6:2153{2175, 2005.

Gunnar Ratsch and Manfred K. Warmuth. Maximizing the margin with boosting. In COLT, pages 334{350, 2002.

Gunnar Ratsch, Sebastian Mika, and Manfred K. Warmuth. On the convergence of leveraging. In NIPS, pages 487{494, 2001.

Sidney I. Resnick. A Probability Path. Birkhauser, 5 edition, 1999.

R. Tyrrell Rockafellar. Convex Analysis Princeton University Press, 1970.

Walter Rudin. Functional Analysis. McGraw-Hill Book Company, 1973.

Robert E. Schapire. The strength of weak learnability. Machine Learning, 5:197{227, July 1990.

Robert E. Schapire. The convergence rate of AdaBoost. In COLT, 2010.

Robert E. Schapire and Yoav Freund. Boosting: Foundations and Algorithms. MIT Press, 2012.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using con dence-rated predictions. Machine Learning, 37(3):297{336, 1999.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the e ectiveness of voting methods. In ICML , pages 322{330, 1997.

Shai Shalev-Shwartz. Introduction to Machine Learning, Course Notes. 2009.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and e cient boosting algorithms. In COLT , pages 311{322, 2008.

Matus Telgarsky. Margins, shrinkage, and boosting. In ICML , 2013.

Lieven Vandenberghe. Ee 236c: Subgradient method. 2013.

Manfred K. Warmuth, Karen A. Glocer, and Gunnar Rätsch. Boosting algorithms for maximizing the soft margin. In NIPS, 2007.

Constantin Zalinescu. Convex analysis in general vector spaces. World scienti c, 2002.

Tong Zhang. Statistical behavior and consistency of classi cation methods based on convex risk minimization. The Annals of Statistics, 32:56{85, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. The Annals of Statistics, 33:1538{1579, 2005.

Index