Uniform Convergence of Rank-weighted Learning

Justin Khim [1] **Liu Leqi** [1] **Adarsh Prasad** [1] **Pradeep Ravikumar** [1]

Abstract

The decision-theoretic foundations of classical machine learning models have largely focused on estimating model parameters that minimize the expectation of a given loss function. However, as machine learning models are deployed in varied contexts, such as in high-stakes decision-making and societal settings, it is clear that these models are not just evaluated by their average performances. In this work, we propose and study a novel notion of L-Risk based on the classical idea of rank-weighted learning. These L-Risks, induced by rank-dependent weighting functions with bounded variation, is a unification of popular risk measures such as conditional value-at-risk and those defined by cumulative prospect theory. We give uniform convergence bounds of this broad class of risk measures and study their consequences on a logistic regression example.

1. Introduction

The statistical decision-theoretic foundations of modern machine learning have largely focused on solving tasks by minimizing the expectation of some loss function. This ensures that the resulting models have high average case performance. However, as machine learning models are deployed along side humans in decision-making, it is clear that they are not just evaluated by their average case performance but also properties like fairness. There has been increasing interest in capturing these additional properties via appropriate modifications of the classical objective of expected loss (Duchi et al., 2019; Garcıa & Fernández, 2015; Sra et al., 2012).

In parallel, there is a long line of work exploring alternatives to expected loss based risk measures in decision-making (Howard & Matheson, 1972), and in reinforcement learning, where percentile based risk measures have been used to quantify the tail-risk of models. A recent line of work borrows classical ideas from behavioral economics for use in machine learning to make models more human-aligned. In particular, Prashanth et al. (2016) have brought ideas from cumulative prospect theory (Tversky & Kahneman, 1992) into reinforcement learning and bandits and Leqi et al. (2019) have used cumulative prospect theory to introduce the notion of human-aligned risk measures.

A common theme in these prior works is the notion of *rank-weighted* risks. The aforementioned risk measures weight each loss by its relative rank, and are based upon the classical rank-dependent expected utility theory (Diecidue & Wakker, 2001). These rank-dependent utilities have also been used in several different contexts in machine learning. For example, such rank-weighted risks have been used to speed up training of deep networks (Jiang et al., 2019). They have also played a key role in designing estimators which are robust to outliers in data. In particular, trimming procedures that simply throw away data with *high* losses have been used to design estimators that are robust to outliers (Daniell, 1920; Bhatia et al., 2015; Lai et al., 2016; Prasad et al., 2018; Lugosi & Mendelson, 2019; Prasad et al., 2019; Shah et al., 2020).

While these rank-based objectives have found widespread use in machine learning, establishing their statistical properties has remained elusive. On the other hand, we have a clear understanding of the generalization properties for average-case risk measures. Loosely speaking, given a collection of models and finite training data, and suppose we choose a model by minimizing average training error, then, we can roughly guarantee on how well this chosen model performs in an average sense. Such guarantees are typically obtained by studying *uniform convergence* of average-case risk measures.

However, as noted before, uniform convergence of rank-weighted risk measures have not been explored in detail. This difficulty comes from the weights being dependent on the *whole* data, thereby, inducing complex dependencies. Hence, existing work on generalization has been on the weaker notion of *pointwise* concentration bounds (Bhat, 2019; Duchi & Namkoong, 2018) or have focused on specific forms of rank-weighted risk measures such as condi-

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA. Correspondence to: Justin Khim <jkhim@cs.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

tional value-at-risk (CVaR) (Duchi & Namkoong, 2018).

Contributions. In this work, we propose the study of rank-weighted risks and prove uniform convergence results. In particular, we propose a new notion of L-Risk in Section 2 that unifies existing rank-dependent risk measures including CVaR and those defined by cumulative prospect theory. In Section 3, we present uniform convergence results, and we observe that the learning rate depends on the weighting function. In particular, when the weighting function is Lipschitz, we recover the standard $O(n^{-1/2})$ convergence rate. Finally, we instantiate our result on logistic regression in Section 4 and empirically study the convergence performance of the L-Risks in Section 6.

2. Background and Problem Setup

In this section, we provide the necessary background on rank-weighted risk minimization, and introduce the notion of bounded variation functions that we consider in this work. Additionally, we provide standard learning-theoretic definitions and notation before moving on to our main results.

2.1. L-Risk Estimation

We assume that there is a joint probability distribution $P(X, Y)$ over the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and our goal is to learn a function $f : \mathcal{X} \mapsto \mathcal{Y}$. In the standard decision-theoretic framework, f^* is chosen among a class of functions $\mathcal{F}$ using a non-negative *loss* function $\ell : \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}_+$.

Classical Risk Estimation. In the traditional setting of risk minimization, the *population* risk of a function f is given by the expectation of the loss function $\ell(f, Z)$ when Z is drawn according to the distribution P:

$$\mathcal{R}(f) = \mathbb{E}_{Z\sim P}[\ell(f, Z)].$$

Given n i.i.d. samples $\{Z_i\}_{i=1}^n$, empirical risk minimization substitutes the empirical risk for the population risk in the risk minimization objective:

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i).$$

L-Risk. As noted in the Section 1, there are many scenarios in machine learning, where we want to evaluate a function by other metrics apart from *average loss*. To this end, we first define the notion of an L-Statistic which dates back to the classical work of (Daniell, 1920).

Definition 1. *Let $X_{(1)} \leq X_{(2)} \ldots \leq X_{(n)}$ be the order statistics of the sample $X_1, X_2, \ldots X_n$. Then, the L-statistic is a linear combination of order statistics,*

$$T_w(\{X_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n w\left(\frac{i}{n}\right) X_{(i)},$$

where $w : [0, 1] \mapsto [0, \infty)$ is a scoring function.

With the above notion of an empirical L-statistic at hand, we define the notion of empirical L-risk for any function f by simply replacing the empirical average of the losses with their corresponding L-statistic.

Definition 2. *The empirical L-risk of f is*

$$\mathcal{LR}_{w,n}(f) = \frac{1}{n}\sum_{i=1}^n w\left(\frac{i}{n}\right) \ell_{(i)}(f),$$

where $\ell_{(1)}(f) \leq \ldots \leq \ell_{(n)}(f)$ are the order statistics of the sample losses $\ell(f, Z_1), \ldots \ell(f, Z_n)$.

Note that the empirical L-risk can be alternatively written in terms of the empirical cumulative distribution function $F_{f,n}$ of the sample losses $\{\ell(f, Z_i)\}_{i=1}^n$:

$$\mathcal{LR}_{w,n} = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) w(F_{f,n}(\ell(f, Z_i))). \quad (1)$$

Accordingly, the population L-risk for any function f can be defined as:

$$\mathcal{LR}_w(f) = \mathbb{E}_{Z\sim P}\left[\ell(f, Z) w(F_f(\ell(f, Z))\right], \quad (2)$$

where $F_f(\cdot)$ is the cumulative distribution function of $\ell(f, Z)$ for Z drawn from P.

2.2. Illustrative Examples of L-Risk

The framework of risk minimization is a central paradigm of statistical estimation and is widely applicable. In this section, we provide illustrative examples that L-risk generalizes classical risk and encompasses several other notions of risk measures. To begin with, observe that simply setting the weighting function as $w(t) = 1$ for all $t \in [0, 1]$, L-Risk minimization corresponds to the classical empirical risk estimation.

Conditional Value-at-Risk (CVaR). As noted in Section 1, in settings where low-probability events have catastrophic losses, using classical risk is inappropriate. Conditional value-at-risk was introduced to handle such tail events and measures the expected loss when conditioned on the event that the loss exceeds a certain threshold. Moreover, CVaR has several desirable properties as a risk measure and in particular, is convex and coherent (Krokhmal et al., 2013). Hence, CVaR is studied across a number of fields such as mathematical finance (Rockafellar et al., 2000), decision making, and more recently machine learning (Duchi et al., 2019). Formally, the CVaR of a function f at a confidence level $1 - \alpha$ is defined as,

$$\mathcal{R}_{\text{CVaR},\alpha}(f) = \mathbb{E}_{Z\sim P}[\ell(f, Z) | \ell(f, Z) \geq \text{VaR}_\alpha(\ell(f, Z))], \quad (3)$$

where $\text{VaR}_\alpha(\ell(f,Z)) = \inf_{x:F_f(x)\geq 1-\alpha} x$ is the value-at-risk. Observe that CVaR is a special case L-Risk in (2) and can be obtained by choosing $w(t) = \frac{1}{\alpha}\mathbf{1}\{t \geq 1-\alpha\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

Human-Aligned Risk (HRM). Cumulative prospect theory (CPT), which is motivated by empirical studies of human decision-making from behavioral economics (Tversky & Kahneman, 1992), has recently been studied in machine learning (Prashanth et al., 2016; Gopalan et al., 2017; Leqi et al., 2019). In particular, Leqi et al. (2019) proposed the following human-aligned risk objective,

$$\mathcal{R}_{\text{HRM},a,b}(f) = \mathbb{E}_{Z\sim P}[\ell(f,Z)w_{\text{a,b}}(F_f(\ell(f,Z))],$$

where $w_{\text{a,b}}(t) = \frac{3-3b}{a^2-a+1}\left(3t^2 - 2(a+1)t + a\right) + 1$.

Trimmed Risk. Trimmed mean is a measure of the central tendency of a distribution and is calculated by discarding samples that are above and below a certain threshold and using the remaining samples to calculate the remaining sample mean. It is known to be more robust to outliers and heavy-tails and is widely used across a variety of disciplines such as finance and aggregating scores in sports. Finally, the trimmed risk of a function f at trimming level α can be defined as,

$$\mathcal{R}_{\text{TRIM},\alpha}(f) = \mathbb{E}_{Z\sim P}[\ell(f,Z)|F_f(\ell(f,Z)) \in [\alpha, 1-\alpha]].$$

The trimmed risk is also a special of L-Risk in (2) and is obtained by setting $w(t) = \frac{1}{1-2\alpha}\mathbf{1}\{\alpha \leq t \leq 1-\alpha\}$.

2.3. Bounded Variation Weighting Functions

Recall from the previous section that the L-risk for any function f depends crucially on the weighting function $w(\cdot)$. Moreover, for popular risk measures such as CVaR and Trimmed Risk, this weighting function is not differentiable and Lipschitz. In this section, we formally introduce a class of weighting functions called *bounded variation* functions, which can be viewed as a strict generalization of Lipschitz functions (Carothers, 2000; Musielak & Orlicz, 1959). More formally, we define p-variation as:

Definition 3. *Let $f : [0,1] \to \mathbb{R}$ be a function. Let $P = \{x_1, \ldots, x_N\} \subset [0,1]$ be a partition of $[0,1]$. Without loss of generality, we assume $x_1 \leq \ldots \leq x_N$. Let $\mathcal{P}$ be the set of all partitions of $[0,1]$, where the size of the partitions may vary between elements of $\mathcal{P}$. The p-variation of f with respect to a partition P is*

$$V_p(f,P) = \left(\sum_{i=1}^{N-1} |f(x_{i+1}) - f(x_i)|^p\right)^{\frac{1}{p}}.$$

The p-variation of f is $V_p(f) = \sup_{P\in\mathcal{P}} V_p(f,P)$.

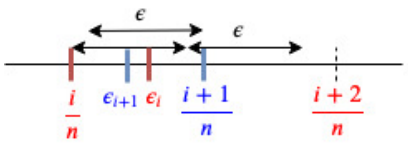


Figure 1. An illustration of the partition argument. Here, we have $\Delta_i = \varepsilon_i$ and $\|\Delta\|_\infty = \varepsilon$. The blue points, $\{(i+1)/n, (i+1)/n + \varepsilon_{i+1}\}$, and the red points, $\{i/n, i/n+\varepsilon_i, (i+2)/n\}$ are in two separate partitions. We construct the partitions this way to ensure that ε_{i+2} and $(i+2)/n$ are not in between i/n and $i/n+\varepsilon_i$ while the order of i/n, ε_i, $(i+1)/n$, and ε_{i+1} do not matter because i/n and $(i+1)/n$ are in separate partitions.

When $p = 1$, the variation $V_p(f) = V_1(f)$ is also called the total variation. Moreover, when the weighting function is λ-Lipschitz, then for all $p \geq 1$, the p-variation is upper bounded by λ.

Table 1 summarizes the bounded variation constants for the aforementioned scoring function. The proofs for these claims can be found in the Appendix. Moving forward, we work with the assumption the weighting function $w(\cdot)$ has bounded variation.

Assumption 1. *The weighting function w has bounded variation $V_p(w)$ for $p = 1, 2$.*

Stability to ℓ_∞ Perturbations. Under Assumption 1, we next present a deterministic bound which controls the stability of bounded-variation functions to ℓ_∞ perturbations.

Lemma 1. *Let $P^* = \{\frac{i}{n}\}_{i=1}^n$ be the n-sized equally spaced partition of the interval $[0,1]$. Then, for any perturbed partition $\widetilde{P} = \{\frac{i}{n} + \Delta_i\}_{i=1}^n$ such that $\sup_i |\Delta_i| = \|\Delta\|_\infty$, we have*

$$\frac{1}{n}\sum_{i=1}^n \left| w\left(\frac{i}{n}\right) - w\left(\frac{i}{n} + \Delta_i\right)\right| \leq \lceil 2n\|\Delta\|_\infty \rceil V_1(w).$$

Proof Sketch. The key idea is that we need to construct sufficiently spaced-out partitions $\mathcal{P}$ so that for all $i \in [n]$, there exists a partition $P \in \mathcal{P}$ such that both i/n and $i/n + \Delta_i$ are in P and no points in between i/n and $i/n + \Delta_i$ are in P. Since for all $i \in [n]$, we know that $i/n + \Delta_i \in (i/n - \|\Delta\|_\infty, i/n + \|\Delta\|_\infty)$, it suffices to have partitions that are $2\|\Delta\|_\infty$ spaced-out. This implies that the total number of partitions we need is at least $\frac{2\|\Delta\|_\infty}{1/n}$. Since w has 1-bounded variation $V_1(w)$, we have reached the desired result. Figure 1 shows the necessity of having spaced-out partitions. □

The above result is a key tool for studying uniform convergence of L-Risks with weighting function w that have

Table 1. Bounded variation for different weight functions.

L-Risk	$w(t)$	$V_1(w)$	$V_2(w)$
$\mathcal{R}$	1	0	0
$\mathcal{R}_{\text{CVaR},\alpha}$	$\frac{1}{\alpha}\mathbf{1}\{1-\alpha \le t\}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$
$\mathcal{R}_{\text{HRM},a,b}$	$\frac{3-3b}{a^2-a+1}\left(3t^2-2(a+1)t+a\right)+1$	$\frac{6(1-b)(2-a)}{a^2-a+1}$	$\frac{6(1-b)(2-a)}{a^2-a+1}$
$\mathcal{R}_{\text{TRIM},\alpha}$	$\frac{1}{1-2\alpha}\mathbf{1}\{\alpha \le t \le 1-\alpha\}$	$\frac{2}{1-2\alpha}$	$\frac{\sqrt{2}}{1-2\alpha}$

bounded variations. It is novel to the best of our knowledge and may be of independent interest.

2.4. Rademacher Complexity, Covering Numbers, and VC Dimension

Finally, we discuss notions of function class complexity that we shall use in our results. We make use of Rademacher complexity, covering numbers, and VC dimension. The reason for all of these is that Rademacher complexity is often an intermediate step but is difficult to analyze. Covering number bounds can be used to help untangle the product structure of the L-Risks, but ultimately to deal with an empirical cumulative distribution function, it is simpler to use VC dimension.

We start with Rademacher complexity. Let $\sigma_1, \dots, \sigma_n$ be i.i.d Rademacher random variables, i.e., random variables that take the values $+1$ and -1 each with probability $1/2$. The empirical Rademacher complexity of a function class $\mathcal{F}$ given a sample $S = Z_1, \dots, Z_n$ is

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i). \tag{4}$$

The Rademacher complexity of $\mathcal{F}$ is then the expectation of the empirical Rademacher complexity with respect to the sample S, i.e.

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_S \hat{\mathfrak{R}}_n(\mathcal{F}).$$

Given a class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, a finite collection of functions $f_1, \dots, f_N$ mapping from $\mathcal{X}$ to $\mathbb{R}$ is called an ε-cover for $\mathcal{F}$ with respect to a semi-norm $\|\cdot\|$ if for every f in $\mathcal{F}$, we have $\min_{j=1,\dots,N} \|f - f_j\| \le \varepsilon$. The ε-covering number of $\mathcal{F}$ with respect to $\|\cdot\|$ is the size of the smallest ε-cover of $\mathcal{F}$ and is denoted by $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$. Note that often bounds are stated in terms of $\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$, which is also called the metric entropy of $\mathcal{F}$. One particular norm of interest is the empirical 2-norm. Given n samples $X_1, \dots, X_n$, define the empirical 2-norm of f to be

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n f(X_i)^2}.$$

Finally, we discuss VC dimension. Let G be a class of functions from $\mathcal{X}$ to some finite set, e.g., $\{0, 1\}$. We define the growth function $\Pi_G : \mathbb{N} \to \mathbb{N}$ by

$$\Pi_G(n) = \max_{x_1,\dots,x_n\in\mathcal{X}} \left|\{(g(x_1), \dots, g(x_n)) : g \in G\}\right|.$$

In words, this is the maximum number of ways that functions in G may classify n points. Now, suppose that functions in G map to a set of two classes, such as $\{0, 1\}$. Then, a set $S = (x_1, \dots, x_n)$ of n points in is said to be shattered by G if there are functions in G realizing all possible label assignments, i.e.,

$$\Pi_G(n) = 2^n = \left|\{(g(x_1), \dots, g(x_n)) : g \in G\}\right|.$$

Finally, the VC-dimension of G, which we denote by $\text{VC}(G)$, is given by

$$\text{VC}(G) = \max\left\{n : \Pi_G(n) = 2^n\right\}. \tag{5}$$

In words, if the VC-dimension of G is V, then there is some set of V points shattered by G. The set that we shall be most interested in using with VC-dimension is

$$\mathcal{L} := \{g_{f,t} : \mathcal{Z} \to \{0, 1\} | g_{f,t}(z) = \mathbf{1}\{\ell(f, z) \le t\} \text{ for } f \in \mathcal{F}, t \in [0, B]\}.$$

Thus, the VC-dimension depends on both the function class $\mathcal{F}$ and the loss function. We note that this is not too large for logistic regression in Lemma 5, and for linear regression in $\mathbb{R}^d$ with squared error loss, the VC dimension is upper bounded by a constant times d (Akama et al., 2010). This VC-dimension plays a key role in the following assumption.

Assumption 2. *Assume that*

$$\sup_{f\in\mathcal{F}} \sup_{x\in[0,B_\ell]} |F_{f,n}(x) - F_f(x)| \le \varepsilon.$$

We make a few remarks related to this assumption. First, it can be thought of as a Glivenko-Cantelli theorem-like assumption, except here we require uniformity over $\mathcal{F}$. Second, our main results relying on variation use this assumption in two places. Third, the assumption can be shown to hold with high probability for $\varepsilon = O(n^{-1/2})$ when $\mathcal{L}$ has bounded VC-dimension via a standard symmetrization

and VC-dimension upper bound argument. We state this sufficient condition as an alternative assumption.

Assumption 2′ *(sufficient condition). The function class $\mathcal{L}$ has bounded VC-dimension.*

Consequently, it is natural to wonder whether this can be relaxed into a statement about the VC-dimension of $\mathcal{F}$. Unfortunately, this result depends on the loss function; so such a general result is unknown to the best of our knowledge. However, we prove it to be the case for logistic regression in Lemma 5, and the proof extends to other widely-used continuous classification losses. For linear regression with squared error loss, this is bounded by a constant times d (Akama et al., 2010). We speculate that $\mathrm{VC}(\mathcal{L})$ is on the order of $\mathrm{VC}(\mathcal{F})$ for reasonable continuous losses.

3. Uniform Convergence Results

In this section, we present our main generalization result in terms of a Rademacher complexity depending on w. Then, we specialize the upper bound using an entropy integral argument in two cases: (a) w with bounded variation and (b) w that is Lipschitz. The former result allows for far more general w, but our proofs lead to slower rates. At best, the rate of $O(n^{-1/4})$ can be achieved by instantiating our bounds, although a more refined argument could possibly improve this. The latter result for Lipschitz w allows for the usual $O(n^{-1/2})$ learning rate.

We use $F_{\mathcal{F}}$ to denote the set $\{F_f\}_{f\in\mathcal{F}}$ and define $\ell(\mathcal{F})\cdot w(F_{\mathcal{F}})$ to be the set $\{\ell_f w(F_f) \mid f\in\mathcal{F}\}$. We have the following bound.

Theorem 1. *Let $\mathcal{F}$ be a set of predictors and the loss function ℓ take values in $[0, B_\ell]$. By Assumption 1, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\sup_{f\in\mathcal{F}}|\mathcal{LR}_{w,n}(f)-\mathcal{LR}_w(f)|\\&\quad\le 2\widehat{\mathfrak{R}}_n(\ell(\mathcal{F})\cdot w(F_{\mathcal{F}}))+4CB_\ell V_1(w)\widehat{\mathfrak{R}}_n(\mathcal{L})\\&\qquad+6B_\ell V_1(w)\sqrt{\frac{\log\frac{4}{\delta}}{2n}}+3\sqrt{\frac{\log\frac{4}{\delta}}{2n}}.\end{aligned}$$

In contrast to standard generalization bounds, our result has the terms $\widehat{\mathfrak{R}}_n(\mathcal{L})$ and $\widehat{\mathfrak{R}}_n(\ell(\mathcal{F})\cdot w(F_{\mathcal{F}}))$. Since the former is a Rademacher complexity of indicator variables, we simply use a VC-dimension upper bound. The VC-dimension then needs to be analyzed for particular losses and $\mathcal{F}$. Examples of losses and $\mathcal{F}$ that permit finite VC-dimension include linear regression with squared error loss and arbitrary $\mathcal{F}$ with logistic loss. To analyze the latter empirical Rademacher complexity, we use the standard Dudley entropy integral result in order to obtain covering numbers. From here, we can more easily decompose the covering number into a product of covering numbers.

Lemma 2. *Suppose that the loss function ℓ is λ_ℓ-Lipschitz in $f(X)$ and bounded by B_ℓ. Additionally, assume that w is bounded by B_w. Then, we have*

$$\begin{aligned}&\mathcal{N}(t,\ell(\mathcal{F})\cdot w(F_{\mathcal{F}}),\|\cdot\|_n)\\&\le\mathcal{N}\left(\frac{t}{2B_w\lambda_\ell},\mathcal{F},\|\cdot\|_n\right)\cdot\mathcal{N}\left(\frac{t}{2B_\ell},w(F_{\mathcal{F}}),\|\cdot\|_n\right).\end{aligned}$$

Now, we use separate tools to analyze the covering number of $w(F_{\mathcal{F}})$ for w of bounded variation and Lipschitz w.

3.1. Weight Functions of Bounded Variation

In this section, we consider w of bounded 2-variation $V_2(w)$; we have the following lemma.

Lemma 3. *Let $\mathcal{C}(\varepsilon', F_{\mathcal{F}})$ be a ε' cover of $F_{\mathcal{F}}$ in $\|\cdot\|_\infty$. Suppose that Assumption 2 holds with ε from the statement of the assumption. Then the set $w(\mathcal{C}(\varepsilon', F_{\mathcal{F}}))$ is a $V_2(w)\sqrt{3(\varepsilon+\varepsilon')}$-cover for $w(F_{\mathcal{F}})$ in $\|\cdot\|_n$.*

Lemma 3 is one of our key technical and conceptual contributions. The key contribution is that we can bound a covering number even though w may be discontinuous, and this relies on constructing a number of partitions that may be upper bounded by total variation as mentioned previously. The shortcoming of this result is also clear. Since ε from Assumption 2 is of order $n^{-1/2}$, the cover is only at the resolution $\varepsilon^{1/2}=n^{-1/4}$.

Corollary 1. *Let $\mathcal{F}$ be a set of predictors, the loss function ℓ take values in $[0, B_\ell]$ and is λ_ℓ-Lipschitz in $f(X)$. If w is bounded by B_w and the VC-dimension V of $\mathcal{L}$ satisfies $1\le V\le n$, by Assumption 1, there exists a universal constant C such that with probability at least $1-\delta$, we have*

$$\begin{aligned}&\sup_{f\in\mathcal{F}}|\mathcal{LR}_{w,n}(f)-\mathcal{LR}_w(f)|\\&\le\inf_{\eta\ge0}\left\{8\eta+\frac{24}{\sqrt{n}}\int_\eta^{2B_\ell B_w}\left(\log\mathcal{N}\left(\frac{t}{2B_w\lambda_\ell},\mathcal{F},\|\cdot\|_n\right)\right.\right.\\&\left.\left.+\log\mathcal{N}\left(\frac{t^2}{12B_\ell^2V_2^2(w)}-\varepsilon,F_{\mathcal{F}},\|\cdot\|_\infty\right)\right)^{1/2}dt\right\}\\&+4CB_\ell V_1(w)\sqrt{\frac{V}{n}}+6B_\ell V_1(w)\sqrt{\frac{\log\frac{4}{\delta}}{2n}}+3\sqrt{\frac{\log\frac{4}{\delta}}{2n}},\end{aligned}$$

where $\varepsilon=2C\sqrt{\frac{V}{n}}+3\sqrt{\frac{\log\frac{4}{\delta}}{2n}}$.

3.2. Lipschitz Weight Functions

In this section, we consider the far simpler case when w is λ_w-Lipschitz. Again, we start with a bound for covering numbers.

Lemma 4. *When w is λ_w-Lipschitz, we have the covering number bound*

$$\mathcal{N}\left(t, w(F_{\mathcal{F}}), \|\cdot\|_n\right) \leq \mathcal{N}\left(t, F_{\mathcal{F}}, \|\cdot\|_\infty\right).$$

This leads naturally to a uniform convergence bound.

Corollary 2. *Let $\mathcal{F}$ be a set of predictors, the loss function ℓ take values in $[0, B_\ell]$ and is λ_ℓ-Lipschitz in $f(X)$. If w is bounded by B_w and is λ_ℓ-Lipschitz and the VC-dimension V of $\mathcal{L}$ satisfies $1 \leq V \leq n$, there exists a universal constant C such that with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\mathcal{LR}_{w,n}(f) - \mathcal{LR}_w(f)| \\
&\leq \inf_{\eta\geq 0}\left\{8\eta + \frac{24}{\sqrt{n}}\int_\eta^{2B_\ell B_w}\left(\log\mathcal{N}\left(\frac{t}{2B_w\lambda_\ell}, \mathcal{F}, \|\cdot\|_n\right)\right.\right. \\
&\left.\left. + \log\mathcal{N}\left(\frac{t}{2B_\lambda\lambda_w}, F_{\mathcal{F}}, \|\cdot\|_\infty\right)\right)^{1/2} dt\right\} \\
&+ 4CB_\ell V_1(w)\sqrt{\frac{V}{n}} + 6B_\ell V_1(w)\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + 3\sqrt{\frac{\log\frac{4}{\delta}}{2n}}.
\end{aligned}$$

Note the key difference between Corollary 1 and Corollary 2. The presence of $t^2 - \varepsilon$ in the former ensures that η needs to be non-zero, and further, $\eta \geq \sqrt{\varepsilon}$. This leads to the slow rate of convergence. In the latter corollary, this is not a problem; thus, we obtain the standard rate of convergence.

4. Logistic Regression Example

We consider a basic example for logistic regression. To instantiate the bounds, we need two things: (1) a bound on the VC dimension of $\mathcal{L}$ and (2) a covering number bound on $F_{\mathcal{F}}$, the distributions of losses over the predictor class $\mathcal{F}$.

Thus, we specify a distribution for (X, Y). Let $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ denote the $(d-1)$-dimensional sphere in $\mathbb{R}^d$, and let $B(d) = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ denote the unit ball of radius 1 in $\mathbb{R}^d$. Let $\theta_* \sim \text{Uniform}(S^{d-1})$ be the true regression vector. Suppose that $X \sim \text{Uniform}(B(d))$ and that Y takes the value $+1$ with probability $p = (1+\theta_*^\intercal X)/2$ and -1 otherwise.

Next, we restrict the class of regressors, $\mathcal{F}$. We set

$$\mathcal{F} = \{f : f(x) = \theta^\intercal x \text{ for } \theta \in \mathbb{R}^d \text{ s.t. } r_1 \leq \|\theta\| \leq r_2\}, \tag{6}$$

where $r_1, r_2 > 0$ are fixed constants. We may refer directly to the parametrizations θ for simplicity. Note that bounding the θ away from the zero vector is necessary because as θ approaches the zero vector, the distribution of losses approaches a step function, making a cover impossible. Finally, recall that the logistic loss is $\ell(f, Z) = \log\left(1 + e^{-Yf(x)}\right)$. Now, we start with our first lemma on VC-dimension.

Lemma 5. *Let $\mathcal{F}$ be the class of linear functions $f(x) = \theta^\intercal x + b$ for x in $\mathbb{R}^d$. Then, the VC-dimension of $\mathcal{L}$ for the logistic loss satisfies*

$$VC(\mathcal{L}) \leq d + 2.$$

This type of bound is not particularly surprising, and other bounds on the VC-dimension of $\mathcal{L}$ in terms of the VC-dimension of $\mathcal{F}$ are known for simple losses. Next, we consider the bound on the covering numbers of the losses.

Lemma 6. *Let (X, Y) have the distribution defined above, and let $\mathcal{F}$ be as in equation (6). Then, we have the ε-entropy bound*

$$\log\mathcal{N}\left(\varepsilon, F_{\mathcal{F}}, \|\cdot\|_\infty\right) \leq \frac{C(\mathcal{F})\sqrt{d}}{\varepsilon}.$$

To the best of our knowledge, such bounds on the set of cumulative distribution functions are novel. Finally, we may apply these lemmas to obtain the following corollaries for weight functions of bounded variation and Lipschitz weight functions. We start with the former.

Corollary 3. *Let (X, Y) have the distribution defined above. Let $\mathcal{F}$ be as defined in equation (6). Suppose that the logistic loss is bounded by B_ℓ on $B(d) \times \{+1, -1\}$ and is λ_ℓ-Lipschitz with respect to $f(x)$ for all f in $\mathcal{F}$. Suppose that w has bounded first and second variations $V_1(w)$ and $V_2(w)$. Then with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\mathcal{LR}_{w,n}(f) - \mathcal{LR}_w(f)| \\
&\qquad \leq C(\mathcal{F}, d, \delta, B_\ell, B_w, V_1(w), V_2(w))n^{-\frac{1}{4}}.
\end{aligned}$$

Next, we consider Lipschitz weight functions.

Corollary 4. *Let (X, Y) have the distribution defined above. Let $\mathcal{F}$ be as defined in equation (6). Suppose that the logistic loss is bounded by B_ℓ on $B(d) \times \{+1, -1\}$ and is λ_ℓ-Lipschitz with respect to $f(x)$ for all f in $\mathcal{F}$. Finally, suppose that w is λ_w-Lipschitz, and let C be some universal constant. Then with probability at least $1-\delta$, we have*

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\mathcal{LR}_{w,n}(f) - \mathcal{LR}_w(f)| \\
&\qquad \leq C(\mathcal{F}, d, \delta, B_\ell, B_w, \lambda_\ell)n^{-\frac{1}{2}}.
\end{aligned}$$

Here, we obtain the desired rates of convergence showing that the generalization bound may be instantiated and lead to quantifiable learning rates.

5. Proofs

In this section, we prove our main theorem (Theorem 1). Lemmas are proved in the Appendix. The first key is to choose the correct decomposition to analyze the main estimator. This is a necessary step because we cannot directly apply standard generalization results to obtain the bound for rank-weighted estimators since the terms $\{F_{f,n}(\ell_{f,i})\}_{i=1}^n$ have used the samples $\{\ell(f, Z_i)\}_{i=1}^n$ twice, and so there is dependence across samples. Our goal is to use symmetrization with the main error.

Lemma 7. *We have the inequality*

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}|\mathcal{LR}_{w,n}(f)-\mathcal{LR}_w(f)|>t+t'\bigg)\\
&\le\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}\bigg|\mathcal{LR}_{w,n}(f)-\frac{1}{n}\sum_{i=1}^n\ell_{f,i}w(F_f(\ell_{f,i}))\bigg|>t\bigg)\\
&+\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\ell_{f,i}w(F_f(\ell_{f,i}))-\mathcal{LR}_w(f)\bigg|>t'\bigg).
\end{aligned}$$

The proof of the lemma is immediate, but it sets the stage to analyze the risk. Next, we present a lemma to deal with the first term on the right hand side of Lemma 7.

Lemma 8. *For some universal constant C, we have*

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}\bigg|\mathcal{LR}_{w,n}(f)-\frac{1}{n}\sum_{i=1}^n\ell_{f,i}w(F_f(\ell_{f,i}))\bigg|\\
&\qquad>4B_\ell V_1(w)\widehat{\mathfrak{R}}_n(\mathcal{L})+6B_\ell V_1(w)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\bigg)\le\delta.
\end{aligned}$$

The proof of this lemma is given in the Appendix due to space concerns. However, note that this is the second lemma that makes use of partitions towards using first bounded variations in its proof. To deal with the second term of Lemma 7, we use a standard Rademacher complexity bound that we provide in the Appendix, which yields

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\ell_{f,i}w(F_f(\ell_{f,i}))-\mathcal{LR}_w(f)\bigg|\\
&\qquad>2\widehat{\mathfrak{R}}_n(\ell(\mathcal{F})\cdot w(F_{\mathcal{F}}))+3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\bigg)\le\delta.
\end{aligned}$$

6. Experiments

Following our theoretical analysis, we test our results on logistic regression and linear regression. We describe our setups in Section 6.1, the metrics as well as the optimization methods we have used in Section 6.2, and finally discuss our results in Sections 6.3.

6.1. Setups

In the logistic regression setup, the features are drawn from $X\sim\text{Uniform}(B(d))$ where $B(d)$ is the d-dimension ball. The labels are sampled from a distribution over $\{-1,1\}$ where Y takes the value $+1$ with probability $(1+X^\top\theta^*)/2$. where $\theta^*=(1,0,0,\cdots,0)\in\mathbb{R}^d$. We use the logistic loss $\ell(\theta;(X,Y))=\log\left(1+\exp(-YX^\top\theta)\right)$.

In the linear regression experiment, we draw our covariates from a Gaussian $X\sim\mathcal{N}(0,\mathbf{I}_d)$ in $\mathbb{R}^d$. The noise distribution is fixed as $\varepsilon\sim\mathcal{N}(0,0.01)$. We draw our response variable Y as, $Y=X^\top\theta^*+\varepsilon$ where $\theta^*=(1,1,\cdots,1)\in\mathbb{R}^d$. We fix the squared error $\ell(\theta;(X,Y))=\frac{1}{2}(Y-X^\top\theta)^2$ as our loss function.

6.2. Metrics and Methods

For each setting, we look at two metrics: (a) approximate uniform error; and (b) training and testing performance of the empirical minimizer of different L-Risks. We approximate the uniform error when $d=1$ in the following manner: for logistic regression, we pick the interval $(-1.5,1.5)$ to be our parameter space Θ and construct a grid of size 200 where each grid point represents a model θ. The true L-Risk for each θ is then approximated by the empirical L-Risk using $20,000$ samples. The empirical L-Risk is calculated using sample size n ranging from 20 to $22,100$. Finally, the approximated uniform error at each sample size n is obtained by taking the maximum over the difference of the empirical and true L-Risk for all θ. In the linear regression setup, the parameter space is chosen to be $(-15,15)$ with size of 500.

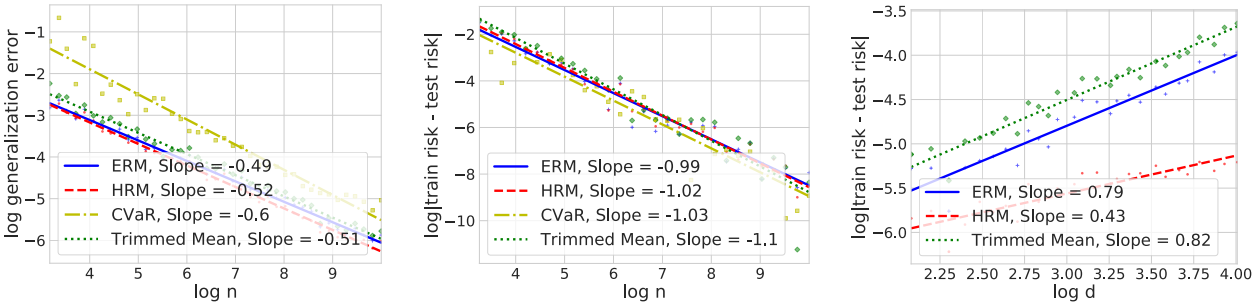
Te explore the training and testing performance of the empirical minimizer of different L-Risks, we perform the following iterative optimization procedure to obtain a heuristic minimizer:

$$\theta^{t+1}=\theta^t-\frac{\gamma_t}{n}\sum_{i=1}^n w_i^t\nabla_\theta\ell(\theta^t;Z_i),\tag{7}$$

for all $t\in\{0,\cdots,T-1\}$, where $w_i^t=w(F_n(\ell(\theta^t;Z_i)))$ is the rank-dependent weighting and γ_t is the learning rate. Although in general, the empirical L-Risks are non-convex with respect to the model parameter θ, there are special cases like CVaR, which has a convex dual form that is easy to optimize (Rockafellar et al., 2000):

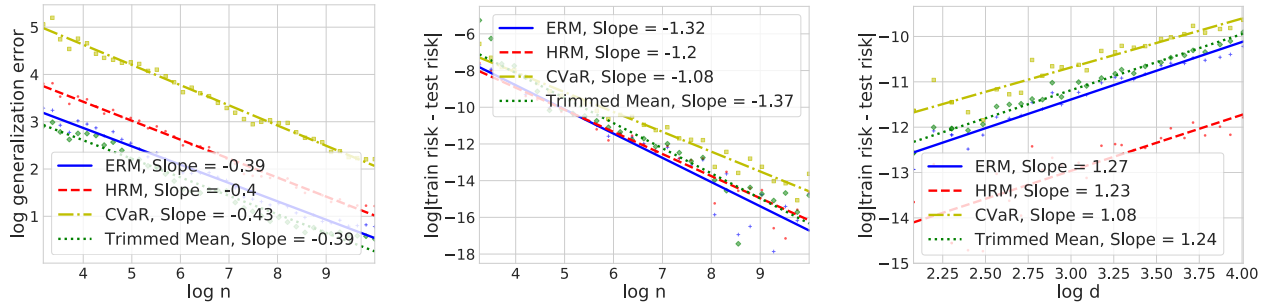
$$\mathcal{R}_{\text{CVaR},\alpha}(f)=\inf_{\eta\in\mathbb{R}}\left\{\eta+\frac{1}{\alpha}\mathbb{E}[\max(\ell_f-\eta,0)]\right\}.\tag{8}$$

In such cases, we compare the heuristic method with the procedure that first optimizes the L-Risk with respect to η and then with respect to θ at each iteration.



(a) log(uniform error) v.s. $\log n$ (b) $d = 5$; n from 20 to 22, 100 (c) $n = 2000$; d from 7 to 55

Figure 2. Experimental results for logistic regression are averaged over 20 seeds. Figure 2a is the log-log plot of the approximated uniform error with respect to different sample sizes. Figure 2b and Figure 2c are log-log plots of the performance of the empirical minimizer of different L-Risks with respect to sample size and sample dimensions. For HRM, we have chosen $a = 0.6$ and $b = 0.4$. For CVaR and trimmed mean, α is set to be 0.1. The results for logistic regression suggests similar convergence behavior across different L-Risks.



(a) log(uniform error) v.s. $\log n$ (b) $d = 5$; n from 20 to 22, 100 (c) $n = 2000$; d from 7 to 55

Figure 3. Experimental results for linear regression are averaged over 20 seeds. Figure 3a is the log-log plot of the approximated uniform error with respect to different sample sizes. Figure 3b and Figure 3c are log-log plots of the performance of the empirical minimizer of different L-Risks with respect to sample size and sample dimensions. For HRM, we have chosen $a = 0.3$ and $b = 0.4$. For CVaR and trimmed mean, α is set to be 0.1. The results for linear regression suggests similar convergence behavior across different L-Risks.

6.3. Results

Our results for logistic regression are given in Figure 2, and our results for linear regression are given in Figure 3. Figure 2a and Figure 3a are the log-log plots of the approximated uniform error with respect to the sample size in the two experimental settings. We fit a line for the data of each L-Risk to examine the rate of convergence of these approximated uniform errors. In both the logistic regression and linear regression experiments, we see that rank-weighting does not have much effect on the convergence rate. In the logistic regression case, the $n^{-1/2}$ convergence rate of HRM matches the expectation since its weight function is Lipschitz. The convergence rate for CVaR and trimmed mean suggests that there might be tighter uniform convergence results for weight function with bounded variation.

Figure 2b, 2c, 3b and 3c show the behavior of the L-Risk of the empirical minimizers. Empirically, we observe that the convergence performance across different L-Risks are similar. Similar to the log-log plots, this may suggest that there are faster rates of convergence than the ones we have obtained in Section 3, especially for the weighting functions that are not Lipschitz but have bounded variation.

Finally, we make a remark about the optimization of CVaR by its rank-weighted formulation versus its convex dual. As shown in Appendix G, we observe that when optimizing CVaR, our heuristic method given in equation (7) improves more quickly than the convex dual form of CVaR equation (8). However, we do note that the dual form achieves a marginally better solution in terms of training error. Practically, the comparable performance of the rank-weighted optimization perspective and the provably convex perspective in the special case where we can compare the two is encouraging for other weight functions. The code of the experiments can be found at `https://bit.ly/2YzwRkJ`.

7. Discussion

In this work, we study uniform convergence of rank-weighted risks, which we define as L-Risk. Different from classical population risks, L-Risk utilizes the ranking of the losses. There are a number of future directions in studying L-Risk. One direction, suggested by our experimental results in Section 6, is to obtain faster rates of convergence for L-Risk with weighting functions of bounded variations. In addition to that, it is also of interest to obtain a lower bound of the convergence rate for L-Risk with non-Lipschitz weighting functions. Finally, since L-Risks are in general non-convex with respect to the model parameters, in order to use it to learn machine learning models, we need to investigate on algorithms for optimizing them. In particular, it would be useful to better understand simple optimization methods for arbitrary rank-weighting functions, such as the iterative approach we have used in the experiments.

Acknowledgements

We acknowledge the support of NSF via IIS-1909816, and ONR via N000141812861.

References

Akama, Y., Irie, K., Kawamura, A., and Uwano, Y. VC dimensions of principal component analysis. *Discrete & Computational Geometry*, 44(3):589–598, 2010.

Bhat, S. P. Improved concentration bounds for conditional value-at-risk and cumulative prospect theory using wasserstein distance. *arXiv preprint arXiv:1902.10709*, 2019.

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pp. 721–729, 2015.

Carothers, N. L. *Real Analysis*. Cambridge University Press, 2000.

Daniell, P. Observations weighted according to order. *American Journal of Mathematics*, 42(4):222–236, 1920.

Diecidue, E. and Wakker, P. P. On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty*, 23(3): 281–298, 2001.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. *preprint*, 2019.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Gautschi, W. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1-4):77–81, 1959.

Gopalan, A., Prashanth, L., Fu, M., and Marcus, S. Weighted bandits or: How bandits learn distorted values that are not expected. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Howard, R. A. and Matheson, J. E. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

Jiang, A. H., Wong, D. L.-K., Zhou, G., Andersen, D. G., Dean, J., Ganger, G. R., Joshi, G., Kaminksy, M., Kozuch, M., Lipton, Z. C., et al. Accelerating deep learning by focusing on the biggest losers. *arXiv preprint arXiv:1910.00762*, 2019.

Krokhmal, P., Zabarankin, M., and Uryasev, S. Modeling and optimization of risk. In *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part II*, pp. 555–600. World Scientific, 2013.

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pp. 665–674. IEEE, 2016.

Leqi, L., Prasad, A., and Ravikumar, P. On human-aligned risk minimization. In *Advances in Neural Information Processing Systems*, 2019.

Lugosi, G. and Mendelson, S. Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391*, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Musielak, J. and Orlicz, W. On generalized variations (i). *Studia Mathematica*, 18(1):11–41, 1959.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Prasad, A., Balakrishnan, S., and Ravikumar, P. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

Prashanth, L., Jie, C., Fu, M., Marcus, S., and Szepesvári, C. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pp. 1406–1415, 2016.

Qi, F. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010:1–84, 2010.

Rockafellar, R. T., Uryasev, S., et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Shah, V., Wu, X., and Sanghavi, S. Choosing the sample with lowest loss makes sgd robust. *arXiv preprint arXiv:2001.03316*, 2020.

Sra, S., Nowozin, S., and Wright, S. J. *Optimization for machine learning*. Mit Press, 2012.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Wainwright, M. J. *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

A. Organization

In this section, we briefly outline the contents and organization of the appendices. In Appendix B, we prove the main lemmas, theorem, and corollaries. The key contribution of this section are the proofs using partitions and variations. In Appendix C, we consider the main proofs for the logistic regression example. In Appendix D, we state and prove auxiliary lemmas. In Appendix E, we state a number of standard results from learning theory. In Appendix F, we prove the basic variation bounds presented in Table 1. In Appendix G, we provide an additional plot on training in our numerical experiments.

B. Main Proofs

In this section, we prove our lemmas. Additionally, we prove the primary corollaries based on weight functions of bounded variation and Lipschitz weight functions.

Proof of Lemma 1. Our goal is to introduce an integer $N \geq 2n\|\Delta\|_\infty$ partitions on $[0,1]$. The reason why we want this is that for any $i \in [n]$, we have $i/n + \Delta_i \in [i/n - \|\Delta\|_\infty, i/n + \|\Delta\|_\infty]$. Thus, for each partition we pick, we should ensure that the points are spread out enough so that for some i, no points will be in between i/n and $i/n + \Delta_i$. In order to do so, we need at least $\frac{2\|\Delta\|_\infty}{1/n}$ partitions since we need separate partitions to cover the points within an interval of size $2\|\Delta\|_\infty$.

For $i = 1, \ldots, N$, consider the set

$$F_i = \bigcup_{j \in J_i} \left\{ \frac{i+jN}{n}, \frac{i+jN}{n} + \Delta_{i+jN} \right\},$$

where $J_i = \{j \geq 0 : i + jN \leq n\}$ is an integer set. Through our way of constructing the partition, we have that

$$\frac{i+(j+1)N}{n} - \frac{i+jN}{n} = \frac{N}{n} = 2\|\Delta\|_\infty.$$

The upshot to this is that if we write $F_i = \{x_{i,j}\}_{j=1}^{2|J_i|}$ such that $x_{i,1} \leq \ldots \leq x_{i,2|J_i|}$, then we have

$$\sum_{i=1}^{n} \left| w\left(\frac{i}{n}\right) - w\left(\frac{i}{n} + \Delta_i\right) \right| \leq \sum_{i=1}^{N} \sum_{j=1}^{2|J_i|-1} |w(x_{i,j+1}) - w(x_{i,j})| \leq \lceil 2n\|\Delta\|_\infty \rceil V_1(w). \tag{9}$$

The reason for the first inequality in equation (9) is that each summand on the left hand side appears on the right hand side, and all additional summands are non-negative. □

Proof of Lemma 8. We have

$$\begin{aligned} P &:= \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell_{f,i} w(F_{f,n}(\ell_{f,i})) - \frac{1}{n} \sum_{i=1}^{n} \ell_{f,i} w(F_f(\ell_{f,i})) \right| > t \right) \\ &\leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |\ell_{f,i}| |w(F_{f,n}(\ell_{f,i})) - w(F_f(\ell_{f,i}))| > t \right) \\ &\leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |w(F_{f,n}(\ell_{f,i})) - w(F_f(\ell_{f,i}))| > \frac{t}{B_\ell} \right) \end{aligned}$$

We start by analyzing the term $A_1 := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} |w(F_{f,n}(\ell_{f,i})) - w(F_f(\ell_{f,i}))|$. Using Assumption 2, let

$$\varepsilon = \sup_{f \in \mathcal{F}} \sup_{x \in [0, B_\ell]} |F_{f,n}(x) - F_f(x)|.$$

Reorder the summands in A_1 so that $w(F_{f,n}(\ell_{f,[i]})) = w(i/n)$. Then, we have

$$A_1 \leq \sup_{\varepsilon_i : |\varepsilon_i| \leq \varepsilon} \sum_{i=1}^{n} \left| w\left(\frac{i}{n}\right) - w\left(\frac{i}{n} + \varepsilon_i\right) \right|.$$

Now, our goal is to introduce an integer $N \geq 2n\varepsilon$ partitions on $[0,1]$. The reason why we want this is that for any $i \in [n]$, $i/n + \varepsilon_i \in [i/n - \varepsilon, i/n + \varepsilon]$. Thus, for each partition we pick, we should ensure that the points are spread out enough so that for some i, no points will be in between i/n and $i/n + \varepsilon_i$. In order to do so, we need at least $\frac{2\varepsilon}{1/n}$ partitions since we need separate partitions to cover the points for each interval of size 2ε.

For $i = 1, \ldots, N$, consider the set

$$F_i = \bigcup_{j \in J_i} \left\{ \frac{i + jN}{n}, \frac{i + jN}{n} + \varepsilon_{i+jN} \right\},$$

where $J_i = \{j \geq 0 : i + jN \leq n\}$ is an integer set. Through our way of constructing the partition, we have that

$$\frac{i + (j+1)N}{n} - \frac{i + jN}{n} = \frac{N}{n} = 2\varepsilon.$$

The upshot to this is that if we write $F_i = \{x_{i,j}\}_{j=1}^{2|J_i|}$ such that $x_{i,1} \leq \ldots \leq x_{i,2|J_i|}$, then we have

$$\sum_{i=1}^{n} \left| w\left(\frac{i}{n}\right) - w\left(\frac{i}{n} + \varepsilon_i\right) \right| \leq \sum_{i=1}^{N} \sum_{j=1}^{2|J_i|-1} \left| w(x_{i,j+1}) - w(x_{i,j}) \right|. \tag{10}$$

The reason for equation (10) is that each summand on the left hand side appears on the right hand side, and all additional summands are non-negative. In particular, we know that for all $k \in \{i + \lfloor i + 2jn\varepsilon \rfloor\}_{j \in J_i}$, k/n and $k/n + \varepsilon_k$ are in F_i, and there are no points in F_i that are in between k/n and $k/n + \varepsilon_k$.

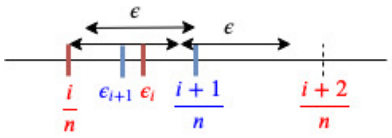


Figure 4. An illustration of constructing the partitions. There are two partitions, blue and red, in this example. Our goal is to construct partitions so that for all $i \in [n]$, i/n and $i/n + \varepsilon_i$ is in the one of the partitions F and there is no points in between in F. Since we have chosen 2ε as the gap between F_{ij} and $F_{i,j+1}$ and $|\varepsilon_i| \leq \varepsilon$, we have ensured that $\{i/n, i/n + \varepsilon_i\}$ are in the red partition with no points in the red partition in between i/n and $i/n + \varepsilon_i$. Since $(i+2)/n$ and i/n are 2ε apart, they can be put in the same red partition since $i/n + \varepsilon_i \leq (i+2)/n - \varepsilon \leq (i+2)/n + \varepsilon_{i+2}$.

Now, we can use total variation to see that the right hand side is bounded by $NV(w)$, and so we have $A_1 \leq NV(w)$. Plugging this back into the definition of A_1, we have

$$P \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \sup_{x \in [0, B_\ell]} |F_{f,n}(x) - F_f(x)| > \frac{t}{2B_\ell V(w)} \right).$$

So, our final task is to upper bound this quantity. For this, we apply Lemma 9. Thus, we obtain

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell_{f,i} w(F_{f,n}(\ell_{f,i})) - \frac{1}{n} \sum_{i=1}^{n} \ell_{f,i} w(F_f(\ell_{f,i})) \right| > 4B_\ell V(w) \widehat{\mathfrak{R}}_n(\mathcal{L}) + 6B_\ell V(w) \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \right) \leq \delta.$$

This completes the proof. □

Proof of Lemma 2. Recall the definition of the set $\ell(\mathcal{F}) \cdot w(F_{\mathcal{F}}) = \{\ell_f w(F_f) \mid f \in \mathcal{F}\}$. Let g be an element of the cover of $\mathcal{F}$, h be an element of the cover of $w(F_{\mathcal{F}})$. For any $f \in \mathcal{F}$, we have

$$\begin{aligned}\|\ell_f w(F_f) - \ell_g h\|_n &\leq \|\ell_f w(F_f) - \ell_f h\|_n + \|\ell_f h - \ell_g h\|_n \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^n B_\ell^2 (w(F_{f,i}) - h(x_i)))^2} + \sqrt{\frac{1}{n}\sum_{i=1}^n B_w^2 (\ell_{f,i} - \ell_{g,i})^2} \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^n B_\ell^2 (w(F_{f,i}) - h(x_i))^2} + \sqrt{\frac{1}{n}\sum_{i=1}^n B_w^2 \lambda_\ell^2 (f(x_i) - g(x_i))^2} \\ &= B_\ell \|w(F_f) - h\|_n + B_w \lambda_\ell \|f - g\|_n.\end{aligned}$$

The proof completes by choosing a $t/(2B_w\lambda_\ell)-$cover of $\mathcal{F}$ and a $t/(2B_\ell)-$cover of $w(F_\mathcal{F})$. □

Proof of Lemma 3. Since $\mathcal{C}(\varepsilon', F_\mathcal{F})$ is a cover of $F_\mathcal{F}$, for every f in $\mathcal{F}$, there is an H in $\mathcal{C}(\varepsilon', F_\mathcal{F})$ such that

$$\begin{aligned}\|w(F_f) - w(H)\|_n &= \sqrt{\frac{1}{n}\sum_{i=1}^n (w(F_f(\ell_{f,i})) - w(H(\ell_{f,i})))^2} \\ &\leq \sup_{\varepsilon_i' : |\varepsilon_i'| \leq \varepsilon'} \sqrt{\frac{1}{n}\sum_{i=1}^n (w(F_f(\ell_{f,i})) - w(F_f(\ell_{f,i}) + \varepsilon_i'))^2}.\end{aligned}$$

Reintroducing $F_{f,n}$ and ε using Assumption 2, we obtain

$$\|w(F_f) - w(H)\|_n \leq \sup_{\varepsilon_i : |\varepsilon_i| \leq \varepsilon} \sup_{\varepsilon_i' : |\varepsilon_i'| \leq \varepsilon'} \sqrt{\frac{1}{n}\sum_{i=1}^n (w(F_{f,n}(\ell_{f,i}) + \varepsilon_i) - w(F_{f,n}(\ell_{f,i}) + \varepsilon_i + \varepsilon_i'))^2}.$$

The advantage to the empirical cdf $F_{f,n}$ is that we can reorder the losses so that $F_{f,n}(\ell_{f,[i]}) = i/n$. Thus, we have

$$\|w(F_f) - w(H)\|_n \leq \sup_{\varepsilon_i : |\varepsilon_i| \leq \varepsilon} \sup_{\varepsilon_i' : |\varepsilon_i'| \leq \varepsilon'} \sqrt{\frac{1}{n}\sum_{i=1}^n \left(w\left(\frac{i}{n} + \varepsilon_i\right) - w\left(\frac{i}{n} + \varepsilon_i + \varepsilon_i'\right)\right)^2}.$$

Now, we can proceed with the total variation bounding technique from the proof of Lemma 8. Let $N \geq 2(\varepsilon + \varepsilon')n$. Then, define

$$F_i = \bigcup_{j \in J_i} \left\{ \frac{i + jN}{n} + \varepsilon_{i+jN}, \frac{i + jN}{n} + \varepsilon_{i+jN} + \varepsilon_{i+jN}' \right\}$$

where $J_i = \{j \geq 0 : i + jN \leq n\}$. Let $F_i = \{x_{i,j}\}_{j=1}^{2|J_i|}$ where $x_{i,1} \leq \ldots \leq x_{i,2|J_i|}$. Then, we see that

$$\sum_{i=1}^n \left(w\left(\frac{i}{n} + \varepsilon_i\right) - w\left(\frac{i}{n} + \varepsilon_i + \varepsilon_i'\right)\right)^2 \leq \sum_{i=1}^N \sum_{j=1}^{2|J_i|-1} |w(x_{i,j+1}) - w(x_{i,j})|^2$$

since every summand on the left hand side appears on the right hand side. This final sum is bounded by $NV_2^2(w)$, and so putting everything together, we have

$$\|w(F_f) - w(H)\|_n \leq V_2(w') \sqrt{\frac{N}{n}} \leq V_2(w) \sqrt{3(\varepsilon + \varepsilon')}.$$

□

Proof of Corollary 1. To prove the corollary, we only need to use Theorem 1 and upper bound the Rademacher complexities by the VC-dimension and appropriate entropy integral. For the empirical Rademacher complexity of , we apply the second inequality of Lemma 9. Next, we notice that $\ell(\mathcal{F}) \cdot w(F_{\mathcal{F}})$ is a set of functions all bounded by $B_\ell B_w$. Thus, by Lemma 16 and Lemma 2, we obtain

$$\widehat{\mathfrak{R}}_n(\ell(\mathcal{F}) \cdot w(F_{\mathcal{F}})) \leq \inf_{\varepsilon \geq 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{2B_\ell B_w} \left(\log \mathcal{N}\left(\frac{t}{2B_w \lambda_\ell}, \mathcal{F}, \|\cdot\|_n \right) + \mathcal{N}\left(\frac{t}{2B_\ell}, w(F_{\mathcal{F}}), \|\cdot\|_n \right) \right)^{\frac{1}{2}} dt \right\}.$$

By Lemma 3, we obtain that

$$\mathcal{N}\left(\frac{t}{2B_\ell}, w(F_{\mathcal{F}}), \|\cdot\|_n \right) \leq \mathcal{N}\left(\frac{t^2}{12 B_\ell^2 V_2^2(w)} - \varepsilon, F_{\mathcal{F}}, \|\cdot\|_n \right),$$

where $\varepsilon = \sup_{f \in \mathcal{F}} \sup_{x \in [0, B_\ell]} |F_{f,n}(x) - F_f(x)|$. By Lemma 9, with probability at least $1 - \delta$ for a universal constant C, we have

$$\varepsilon \leq 2C \sqrt{\frac{V}{n}} + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

This completes the proof. □

Proof of Corollary 2. To prove this corollary, we again only need to use Theorem 1, the second inequality of Lemma 9, and an upper bound the other Rademacher complexity. Again using Lemma 16 and Lemma 2, we obtain

$$\widehat{\mathfrak{R}}_n(\ell(\mathcal{F}) \cdot w(F_{\mathcal{F}})) \leq \inf_{\varepsilon \geq 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{2B_\ell B_w} \left(\log \mathcal{N}\left(\frac{t}{2B_w \lambda_\ell}, \mathcal{F}, \|\cdot\|_n \right) + \mathcal{N}\left(\frac{t}{2B_\ell}, w(F_{\mathcal{F}}), \|\cdot\|_n \right) \right)^{\frac{1}{2}} dt \right\}.$$

By Lemma 4, we have

$$\mathcal{N}\left(\frac{t}{2B_\ell}, w(F_{\mathcal{F}}), \|\cdot\|_n \right) \leq \mathcal{N}\left(\frac{t}{2B_\ell \lambda_w}, F_{\mathcal{F}}, \|\cdot\|_\infty \right),$$

and this completes the proof. □

C. Logistic Regression Proofs

In this section, we prove our statements for our logistic regression examples. The two key lemmas introduced in the main paper are a bound on the VC-dimension of $\mathcal{L}$ and a bound on the metric entropy of $F_{\mathcal{F}}$. The latter requires basic bounds on the volume of hyperspherical caps and ratios of the Gamma function, which are contained in the subsequent section. Additionally, in this section, we prove our corollaries for different weight functions.

Proof of Lemma 5. First, we start by examining elements of $\mathcal{L}$. Let $g_{f,t}$ be in $\mathcal{L}$. Then, $g_{f,t}$ has the form

$$g_{f,t}(z) = \mathbf{1}\left\{ \ell(f, z) \leq t \right\} = \mathbf{1}\left\{ \log(1 + \exp(-y f(x))) \leq t \right\}.$$

By basic algebraic manipulations, this is equivalent to

$$g_{f,t}(z) = \mathbf{1}\left\{ -y f(x) \leq \log(\exp(t) - 1) \right\}.$$

Since the choice of t is arbitrary and $m(t) = \log(\exp(t) - 1)$ is a bijection of the real line, we may reparametrize and write

$$g_{f,t}(z) = \mathbf{1}\left\{ -y f(x) \leq t \right\}.$$

Next, we use the fact that f is a linear classifier; so we can write

$$g_{f,t}(z) = \mathbf{1}\left\{ -y(\theta^\intercal x + b) \leq t \right\}.$$

Now, our goal is to use an invariance observation to remove the $(-Y)$ from the beginning of the equation. For this, we again alter $g_{f,t}$, and this time we consider

$$h_{f,t}(z) = \text{sgn}\left((-y)((\theta,t)^\intercal(x,y)+b)\right) = -y\text{sgn}\left((\theta,t)^\intercal(x,y)+b\right).$$

Thus, instead of considering mappings to $\{0,1\}$, we consider mappings to $\{+1,-1\}$. Let $\mathcal{Z}_D = \{z_1,\ldots,z_D\}$ be arbitrary, and let $\mathcal{H}$ be the set of all functions $h : \mathcal{Z}_D \to \{+1,-1\}$. To show that $\mathcal{L}$ has VC dimension bounded by some D, we need to prove that there does not exist any set $\mathcal{Z}_D$ such that, for all h in $\mathcal{H}$, we have $h_{f,t}(z) = h(z)$ on $\mathcal{Z}_D$ for some f and t.

Now, we have the key observation: it suffices to instead show that there does not exist a set $\mathcal{Z}_D$ such that, for all h in $\mathcal{H}$, we have $h_{f,t}(z) = -yh(z)$ on $\mathcal{Z}_D$ for some f and t. The reason for this is simple. If we define

$$-y\mathcal{H} = \{(x,y) \mapsto -yh(x) : h \in \mathcal{H}\},$$

then $-y\mathcal{H}$ and $\mathcal{H}$ are equivalent.

Using this observation, we only need to examine the conditions under which

$$h_{f,t}(z) = -y\text{sgn}((\theta,t)^\intercal(x,y)+b) = -yh(z).$$

Multiplying both sides by $-y$, we need to find the size of the largest set such that we can find f and t satisfying

$$\text{sgn}((\theta,t)^\intercal(x,y)+b) = h(z) \tag{11}$$

for all z in $\mathcal{Z}_D$ for arbitrary h. But, equation (11) is a restricted form of linear classification in $\mathbb{R}^{d+1}$. The VC dimension of linear classifiers over $\mathbb{R}^{d+1}$ is $d+2$; thus, the VC dimension of $\mathcal{L}$ is upper bounded by $d+2$. □

Proof of Lemma 6. Our strategy is to define the cumulative distribution function for a particular classifier, F_θ; to show that this cdf is Lipschitz; and then to use the standard argument for covering a Lipschitz function class. We start with the first task. Let $\text{Vol}(B(d))$ denote the volume of the unit ball in $\mathbb{R}^d$. The cdf of the loss given the classifier θ is

$$\begin{aligned} F_\theta(t) = &\int_{B(d)} \frac{1}{\text{Vol}(B(d))}\left(\frac{1+\theta_*^\intercal x}{2}\right)\mathbf{1}\left\{\ell(\theta,(x,+1)) \le t\right\}dx \\ &+ \int_{B(d)} \frac{1}{\text{Vol}(B(d))}\left(\frac{1-\theta_*^\intercal x}{2}\right)\mathbf{1}\left\{\ell(\theta,(x,-1)) \le t\right\}dx. \end{aligned} \tag{12}$$

Now, we move to the second step of the proof: proving that F_θ is Lipschitz. Suppose that $t \le t'$. Using equation (12), we have

$$\begin{aligned} F_\theta(t') - F_\theta(t) = &\int_{B(d)} \frac{1}{\text{Vol}(B(d))}\left(\frac{1+\theta_*^\intercal x}{2}\right)\mathbf{1}\left\{t < \ell(\theta,(x,+1)) \le t'\right\}dx \\ &+ \int_{B(d)} \frac{1}{\text{Vol}(B(d))}\left(\frac{1-\theta_*^\intercal x}{2}\right)\mathbf{1}\left\{t < \ell(\theta,(x,-1)) \le t'\right\}dx \\ \le &\frac{1}{\text{Vol}(B(d))}\int_{B(d)} \mathbf{1}\left\{t < \ell(\theta,(x,+1)) \le t'\right\}dx \\ &+ \frac{1}{\text{Vol}(B(d))}\int_{B(d)} \mathbf{1}\left\{t < \ell(\theta,(x,-1)) \le t'\right\}dx. \end{aligned}$$

At this point, we need to use the form of the loss to make further progress on the integral. The first indicator is

$$\mathbf{1}\left\{t < \ell(\theta,(x,+1)) \le t'\right\} = \mathbf{1}\left\{t < \log(1+\exp(-\theta^\intercal x)) \le t'\right\} = \mathbf{1}\left\{g(t) \le -\theta^\intercal x \le g(t')\right\}$$

where $g(s) = \log(\exp(s)-1)$. Presently, we need to use the distribution of X to deduce the distribution of $-\theta^\intercal x$. Since X is uniformly distributed on the sphere, by spherical symmetry, we have the distributional equivalence $-\theta^\intercal X = \|\theta\| X_1$, where X_1 is the first component of X. The same analysis shows that $\theta^\intercal X = \|\theta\| X_1$; so the following analysis applies to the second integral as well. The quantity we need to bound is

$$C = \int_{B(d)} \mathbf{1}\left\{\frac{g(t)}{\|\theta\|} < x_1 \le \frac{g(t')}{\|\theta\|}\right\}dx.$$

The integral is a difference in the volume of spherical caps when $g(t)$ and $g(t')$ have the same sign; otherwise the integral spans the equator of the hypersphere. Thus, let $V(h, r)$ be the volume of the spherical cap of height h and radius r.

We consider three cases involiving the values of $g(t)$ and $g(t')$, but we first make a simplifying observation. Since X takes values on the unit sphere and θ is bounded, it suffices to consider t and t' in the interval $[L_*, L^*]$ where $L_* = \log(1 + \exp(-\|\theta\|))$ and $L^* = \log(1 + \exp(\|\theta\|))$. Otherwise, we have $F_\theta(t) = 0$ for $t \leq L_*$ and $F_\theta(t) = 1$ for $t \geq L^*$. Thus, in proving that F_θ is Lipschitz, it suffices to consider the interval $[L_*, L^*]$. Note that $g(L_*) = -\|\theta\|$ and $g(L^*) = \|\theta\|$ as well. Now, we proceed to our cases.

Case 1: $g(L_*) \leq g(t) \leq g(t') \leq 0$. In this case, we have

$$C = V\left(1 + \frac{g(t')}{\|\theta\|}, 1\right) - V\left(1 + \frac{g(t)}{\|\theta\|}, 1\right).$$

Using Lemma 10, we have

$$\begin{aligned} F_\theta(t') - F_\theta(t) &\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\left(\int_{-g(t')/\|\theta\|}^{1}(1-\phi^2)^{\frac{d-1}{2}}d\phi - \int_{-g(t)/\|\theta\|}^{1}(1-\phi^2)^{\frac{d-1}{2}}d\phi\right) \\ &= 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\int_{-g(t')/\|\theta\|}^{-g(t)/\|\theta\|}(1-\phi^2)^{\frac{d-1}{2}}d\phi \\ &\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))\|\theta\|}\left|g(t') - g(t)\right|. \end{aligned}$$

At this point, we consider our other cases, for which we derive similar equations.

Case 2: $0 \leq g(t) \leq g(t') \leq g(L^*)$. In this case, we have

$$C = V\left(1 - \frac{g(t)}{\|\theta\|}, 1\right) - V\left(1 - \frac{g(t)}{\|\theta\|}, 1\right).$$

Using Lemma 10, we have

$$\begin{aligned} F_\theta(t') - F_\theta(t) &\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\left(\int_{g(t)/\|\theta\|}^{1}(1-\phi^2)^{\frac{d-1}{2}}d\phi - \int_{g(t')/\|\theta\|}^{1}(1-\phi^2)^{\frac{d-1}{2}}d\phi\right) \\ &= 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\int_{g(t)/\|\theta\|}^{g(t')/\|\theta\|}(1-\phi^2)^{\frac{d-1}{2}}d\phi \\ &\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))\|\theta\|}|g(t') - g(t)|. \end{aligned}$$

This gives the same inequality as for case 1; now we proceed to the final case.

Case 3: $g(L_*) \leq g(t) \leq 0 \leq g(t') \leq g(L^*)$. Here, we have

$$C = V(1,1) - V\left(1 - \frac{g(t')}{\|\theta\|}, 1\right) + V(1,1) - V\left(1 + \frac{g(t)}{\|\theta\|}, 1\right).$$

Again using Lemma 10, we have

$$\begin{aligned}
F_\theta(t') - F_\theta(t) &\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\left(\int_0^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi - \int_{g(t')/\|\theta\|}^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi\right. \\
&\qquad \left. + \int_0^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi - \int_{-g(t)/\|\theta\|}^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi\right) \\
&= 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\left(\int_0^{g(t')/\|\theta\|} (1-\phi^2)^{\frac{d-1}{2}}d\phi + \int_0^{-g(t)/\|\theta\|} (1-\phi^2)^{\frac{d-1}{2}}d\phi\right) \\
&= 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}\int_{g(t)/\|\theta\|}^{g(t')/\|\theta\|} (1-\phi^2)^{\frac{d-1}{2}}d\phi \\
&\leq 2\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))\|\theta\|}|g(t') - g(t)|.
\end{aligned}$$

Thus, we have established the previous equation in all cases.

Finally, for all of the cases above, we simply use the fact that g is Lipschitz on the interval $[L_*, L^*]$; moreover, there is some worst case Lipschitz constant for g and all parametrizations θ since $\mathcal{F}$ is compact. Using this constant C', we have

$$F_\theta(t') - F_\theta(t) \leq \frac{2}{\|\theta\|}\cdot\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}C'|t'-t| \leq \frac{2C'}{r_1}\cdot\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))}|t'-t|.$$

This proves that F_θ is Lipschitz, and we may pick the same Lipschitz constant for all θ. Next, we take a minor detour to investigate the dependence of the Lipschitz constant on d, which involves analyzing the ratio of ball volumes.

The volume of the unit ball in $\mathbb{R}^d$ is $\pi^{d/2}/\Gamma(d/2+1)$. As a result, we have

$$\frac{\text{Vol}(B(d-1))}{\text{Vol}(B(d))} = \frac{1}{\sqrt{\pi}}\cdot\frac{\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)} \leq \sqrt{\frac{d}{\pi}},$$

where the inequality is due to Lemma 11. Thus, we have

$$F_\theta(t') - F_\theta(t) \leq \frac{2C'}{r_1\sqrt{\pi}}\sqrt{d}|t'-t| = C(\mathcal{F})\sqrt{d}|t'-t|.$$

Finally, we proceed to the third step of the proof: covering $\mathcal{F}$ in the infinity norm. We have shown that each F_θ in $\mathcal{F}$ is a $C(\mathcal{F})\sqrt{d}$-Lipschitz function on $[L_*, L^*]$. From the fact that each F_θ is a cumulative distribution function, we know that they are monotonically increasing, take values between 0 and 1, and start at 0 in that $\inf_t F_\theta(t) = 0$. This set is ε-coverable in the infinity norm by $2^{(L^*-L_*)C(\mathcal{F})\sqrt{d}/\varepsilon}$ functions that are piecewise linear on the grid with spaces in t of size $\varepsilon/(C(\mathcal{F})\sqrt{d})$ and slopes of either 0 or $C(\mathcal{F})\sqrt{d}$ on each segment. As a result, the ε-entropy of $\mathcal{F}$ is

$$\log\mathcal{N}(\varepsilon, F_\mathcal{F}, \|\cdot\|_\infty) \leq \log 2^{\frac{(L^*-L_*)C(\mathcal{F})\sqrt{d}}{\varepsilon}} \leq \frac{(L^*-L_*)C(\mathcal{F})\sqrt{d}}{\varepsilon}.$$

Note that $L^* - L_* \leq 4r_2$, so the dependence of the cover size on $\mathcal{F}$ is at least r_2/r_1. This completes the proof. □

Proof of Corollary 3. From Corollary 1, it suffices to bound the VC dimension of $\mathcal{L}$ for the logistic loss and the first term of the bound. For the former, we simply apply Lemma 5.

Now, we consider the first term of the bound. We ultimately choose $\eta = B_\ell V_2(w)(12\varepsilon)^{1/2}$, but for simplicity, we maintain

the η through our initial calculations. We focus on the entropy integral, which we denote by J. We have

$$\begin{aligned}
J &= \int_{\eta}^{2B_\ell B_w} \left(\log \mathcal{N}\left(\frac{t}{2B_w\lambda_\ell}, \mathcal{F}, \|\cdot\|_n \right) + \log \mathcal{N}\left(\frac{t^2}{12B_\ell^2 V_2^2(w)} - \varepsilon, F_{\mathcal{F}}, \|\cdot\|_\infty \right) \right)^{\frac{1}{2}} dt \\
&\leq \int_{\eta}^{2B_\ell B_w} \left(d \log\left(1 + \frac{4r_2 B_w \lambda_\ell}{t}\right) + \frac{12B_\ell^2 V_2^2(w) C(\mathcal{F})\sqrt{d}}{t^2 - 12B_\ell^2 V_2^2(w)\varepsilon} \right)^{\frac{1}{2}} dt \\
&\leq \int_{\eta}^{2B_\ell B_w} \left(\frac{4dr_2 B_w \lambda_\ell}{t} \right)^{\frac{1}{2}} dt + \int_{\eta}^{2B_\ell B_w} \left(\frac{12B_\ell^2 V_2^2(w) C(\mathcal{F})\sqrt{d}}{t^2 - 12B_\ell^2 V_2^2(w)\varepsilon} \right)^{\frac{1}{2}} dt,
\end{aligned}$$

where the first inequality follows from Lemma 6 and Lemma 12 and the second inequality follows from the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for non-negative x and y. Integrating, we obtain

$$\begin{aligned}
J &\leq 2\sqrt{8dr_2 B_w \lambda_\ell B_\ell B_w} + \log\left(2B_\ell B_w + \sqrt{(2B_\ell B_w)^2 - 12B_\ell^2 V_2^2(w)\varepsilon} \right) \\
&\quad - \log\left(\eta + \sqrt{\eta^2 - 12B_\ell^2 V_2^2(w)\varepsilon} \right) \\
&\leq 2\sqrt{8dr_2 B_w \lambda_\ell B_\ell B_w} + \log\left(4B_\ell B_w\right) - \log(\eta) \\
&\leq C + C' \log n,
\end{aligned}$$

where the final inequality follows since $\eta = \Theta(n^{-1/2})$.

Thus, putting everything together with the bound of Corollary 1, we have

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell_{f,i} w(F_{f,n}(\ell_{f,i})) - \mathbb{E}[\ell_f w(F_f(\ell_f))] \right| \leq C(\mathcal{F}, d, \delta, B_\ell, B_w, V_1(w), V_2(w)) n^{-\frac{1}{4}},$$

since $\eta = \Theta(n^{-1/4})$. This completes the proof. □

Proof of Corollary 4. From Corollary 2, it suffices to analyze the VC dimension of $\mathcal{L}$ for the logistic loss and the first term of the bound. For the former, we simply apply Lemma 5.

Now, we consider the first term in the bound. To simplify matters, we pick $\eta = 0$. Denote the entropy integral by J. Using Lemma 12 and Lemma 6, we have

$$\begin{aligned}
J &= \int_{0}^{2B_\ell B_w} \left(d \log\left(1 + \frac{2r_2}{t}\right) + \frac{2B_\ell \lambda_w C(\mathcal{F})\sqrt{d}}{t} \right)^{\frac{1}{2}} dt \\
&\leq \sqrt{2dr_2 + 2B_\lambda \lambda_w C(\mathcal{F})\sqrt{d}} \int_{0}^{2B_\ell B_w} \frac{1}{\sqrt{t}} dt \\
&= \sqrt{B_\ell B_w (2dr_2 + 2B_\lambda \lambda_w C(\mathcal{F})\sqrt{d})}.
\end{aligned}$$

Plugging this result in completes the proof. □

D. Additional Lemmas

In this section, we prove a few auxiliary lemmas.

Lemma 9. *Let V be the VC-dimension of $\mathcal{L}$. With probability at least $1-\delta$, for a universal constant C, we have*

$$\sup_{f \in \mathcal{F}} \sup_{x \in [0, B_\ell]} |F_{f,n}(x) - F_f(x)| \leq 2\widehat{\mathfrak{R}}_n(\mathcal{L}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \leq 2C\sqrt{\frac{V}{n}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Proof. Our goal is to use symmetrization and a VC-dimension bound. First, let $\ell'_{f,i} = \ell(f, Z'_i)$ be a random variable independent of and identically distributed to $\ell_{f,i}$. Define S' to be the set of $\ell'_{f,1}, \ldots, \ell'_{f,n}$. Then, we can write

$$\sup_{f\in\mathcal{F}}\sup_{x\in[0,B_\ell]}|F_{f,n}(x)-F_f(x)| = \sup_{f\in\mathcal{F}}\sup_{x\in[0,B_\ell]}\frac{1}{n}\left|\sum_{i=1}^n \mathbf{1}\left\{\ell_{f,i}\le x\right\} - \mathbb{E}_{S'}\sum_{i=1}^n \mathbf{1}\left\{\ell'_{f,i}\le x\right\}\right|.$$

Then by Lemma 13, Lemma 16 and Lemma 17, with probability at least $1-\delta$, for a universal constant C, we have

$$\sup_{f\in\mathcal{F}}\sup_{x\in[0,B_\ell]}|F_{f,n}(x)-F_f(x)| \le 2\widehat{\mathfrak{R}}_n(\mathcal{L}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \le 2C\sqrt{\frac{V}{n}} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

This completes the proof. □

Lemma 10. *Let S be a sphere of radius r in $\mathbb{R}^d$. Let $V(h,r)$ be the volume of the spherical cap of height h. Then, the volume of the cap is*

$$V(h,r) = r^d \mathrm{vol}(B(d-1))\int_{\frac{r-h}{r}}^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi.$$

Proof. The proof is a straightforward integration, followed by a change of variables and use of trigonometric identities. We start by observing that we can calculate the volume of the cap by integrating the volume of a $(d-1)$-sphere with radius $r\sin\theta$ over θ from 0 until $r\cos\theta = r-h$. This integral is

$$\begin{aligned} V(h,r) &= \mathrm{vol}(B(d-1))\int_0^{\arccos\left(\frac{r-h}{r}\right)} r^{d-1}\sin^{d-1}(\theta) r\sin(\theta)d\theta \\ &= r^d\mathrm{vol}(B(d-1))\int_0^{\arccos\left(\frac{r-h}{r}\right)}\sin^d\theta d\theta. \end{aligned}$$

Next, we use a change of variables, setting $\theta = \arccos\phi$. This gives

$$\begin{aligned} V(h,r) &= r^d\mathrm{vol}(B(d-1))\int_1^{\frac{r-h}{r}} \sin^d(\arccos\phi)\left(-\frac{1}{\sqrt{1-\phi^2}}\right)d\phi \\ &= r^d\mathrm{vol}(B(d-1))\int_{\frac{r-h}{r}}^1 (1-\phi^2)^{\frac{d-1}{2}}d\phi, \end{aligned}$$

where in the second line we use the trigonometric identity $\sin(\arccos\phi) = (1-\phi^2)^{1/2}$. This completes the proof. □

We also need a result on ratios of the Gamma function (Gautschi, 1959; Qi, 2010).

Lemma 11 (Gautschi's inequality)**.** *For $x>0$ and s in $(0,1)$, we have the inequalities*

$$x^{1-s} \le \frac{\Gamma(x+1)}{\Gamma(x+s)} \le (x+s)^{1-s}.$$

Lemma 12. *Suppose that $\mathcal{F}$ is a class of linear functions $f(x) = \theta^\intercal x$ where $\|\theta\| \le r$. Assume that $X_1, \ldots, X_n$ take values in the unit ball $B(d)$. Then, we have the metric entropy bound*

$$\log\mathcal{N}(t,\mathcal{F},\|\cdot\|_n) \le d\log\left(1+\frac{2r}{t}\right).$$

Proof. Let $f(x) = \theta^\intercal x$ and $g(x) = \eta^\intercal x$ be two functions in $\mathcal{F}$. Observe that

$$\|f - g\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n [(\theta - \eta)^\intercal X_i]^2} \leq \|\theta - \eta\| \sqrt{\frac{1}{n}\sum_{i=1}^n \|X_i\|^2} \leq \|\theta - \eta\|.$$

Thus, it suffices to produce a cover of $\mathcal{F}$ in the Euclidean norm $\|\cdot\|$. By Lemma 18, the metric entropy of a unit ball is bounded above by $d\log(1 + 2/\delta)$. Since $\mathcal{F}$ is contained in a ball of radius r, we have

$$\log \mathcal{N}(t, \mathcal{F}, \|\cdot\|) \leq \log \mathcal{N}\left(\frac{t}{r}, B(d), \|\cdot\|\right) \leq d \log\left(1 + \frac{2r}{t}\right),$$

which gives the desired bound. □

E. Standard Lemmas

In this section, we provide a number of standard lemmas used in our results.

Lemma 13 (Theorem 3.1 of Mohri et al. 2012)**.** *Let G be a family of functions mapping from $\mathcal{Z}$ to $[0, 1]$. Then for $\delta > 0$, with probability at least $1 - \delta$, we have that for all $g \in G$,*

$$\mathbb{E}g(Z) \leq \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\mathfrak{R}_n(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\textit{and} \quad \mathbb{E}g(Z) \leq \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\widehat{\mathfrak{R}}_n(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

We now present a few lemmas towards using VC dimension bounds.

Lemma 14 (Corollary 3.1 of Mohri et al. 2012)**.** *Let $\mathcal{F}$ be a class of functions taking values in $\{+1, -1\}$. Then, we have the following bound*

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log \Pi_{\mathcal{F}}(n)}{n}}.$$

Note that the previous lemma also applies to functions taking values in $\{+1, 0, -1\}$, since the only way this is used is to bound the norm of $(f(z_1), \ldots, f(z_n))$, which is still bounded by $\sqrt{n}$ when allowing for zeros.

Lemma 15 (Corollary 3.3 of Mohri et al. 2012)**.** *Let $\mathcal{F}$ be a function class with VC dimension V. Then for all $n \geq d$, we have*

$$\Pi_{\mathcal{F}}(n) \leq \left(\frac{en}{V}\right)^V.$$

Lemma 16 (Theorem 5.22 and Example 5.24 of Wainwright 2019)**.** *For a class of $b-$uniformly bounded functions $\mathcal{F}$, we have*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leq \inf_{\eta \geq 0}\left\{8\eta + \frac{32}{\sqrt{n}}\int_\eta^{2b} \sqrt{\log \mathcal{N}(t, \mathcal{F}, \|\cdot\|_n)}dt\right\}.$$

The following allows us to bound covering numbers in terms of VC dimension. Theorem 2.6.4 and Theorem 2.6.7 of (Van Der Vaart & Wellner, 1996) provide similar results.

Lemma 17 (Equation 5.49 of Wainwright 2019)**.** *For a class of $b-$uniformly bounded functions $\mathcal{F}$ with VC dimension V there is a universal constants C such that*

$$\mathcal{N}(t, \mathcal{F}, \|\cdot\|_n) \leq CV(16e)^V \left(\frac{b}{t}\right)^{2V}.$$

Lemma 18 (Example 5.9 of Wainwright 2019)**.** *Let $B = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ be a unit ball with respect to norm $\|\cdot\|$. Then, the metric entropy of B in terms of its own norm $\|\cdot\|$ is bounded above and below as*

$$d \log\left(\frac{1}{t}\right) \leq \log \mathcal{N}(t, B, \|\cdot\|) \leq d \log\left(1 + \frac{2}{t}\right).$$

Lemma 19. *Let X be a $d-$dimensional sub-Gaussian random vector with parameter σ^2. With probability $1-\delta$, we have*

$$\|X\|_2 \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{\log\frac{1}{\delta}}.$$

Proof. Let $\mathbb{S}^{d-1}$ denote the set of vectors $v \in \mathbb{R}^d$ such that $\|v\|_2 \leq 1$. Let $\mathcal{C}(1/2, \mathbb{S}^{d-1}, \|\cdot\|_2)$ denote a $1/2-$cover of $\mathbb{S}^{d-1}$ in Euclidean norm. We have that

$$\|X\|_2 \leq \max_{v \in \mathbb{R}^d, \|v\|_2 \leq 1} v^\top X \leq \max_{u \in \mathcal{C}(1/2, \mathbb{S}^{d-1}, \|\cdot\|_2)} u^\top X + \frac{1}{2}\|X\|_2.$$

By definition, we know that $u^\top X$ is a sub-Gaussian random variable with parameter σ^2. By Lemma 18, we have $|T| \leq 5^d$. Then, we have that

$$\mathbb{P}(\|X\|_2 \geq t) \leq \mathbb{P}(2 \max_{u \in T} u^\top X \geq t) \leq |T| \mathbb{P}(u^\top X \geq t/2) \leq 5^d e^{-\frac{t^2}{8\sigma^2}}.$$

□

F. Proofs of Table 1

In this section, we provide a few of the bounds in Table 1. For the usual risk and HRM, we use a general bound for Lipschitz weight functions. For CVaR, we provide a simple proof. The proof for the trimmed mean is similar.

Proof of Bounded Variation of Lipschitz w. Fix a partition $P = \{x_1, \ldots, x_N\}$ such that $x_1 \leq \ldots \leq x_N$. Then, we have

$$V_p(f, P) = \left(\sum_{i=1}^{N-1} |f(x_{i+1}) - f(x_i)|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{N-1} L^p |x_{i+1} - x_i|^p\right)^{\frac{1}{p}} \leq L(x_N - x_1) \leq L.$$

Since this upper bound holds for any partition, we have $V(f) \leq L$, as desired. □

Proof of Bounded Variation of CVaR. Let $P = \{x_1, \ldots, x_N\}$ be a partition such that $x_1 \leq \ldots \leq x_N$. Note that the scoring function of CVaR is $w_{\text{CVaR}}(t) = \frac{1}{\alpha}\mathbf{1}\{t \geq 1-\alpha\}$ for $t \in [0,1]$. There are two cases.

Case 1: we have $P \subset [0, 1-\alpha)$ or $P \subset [1-\alpha, 1]$. In this case, $w_{\text{CVaR}}(x_i) = w_{\text{CVaR}}(x_j)$ for all $i, j = 1, \ldots, N$. Thus, we see $V_p(w_{\text{CVaR}}, P) = 0$.

Case 2: there is an i for which $x_i < 1-\alpha$ and $x_{i+1} \geq 1-\alpha$. Here, we have

$$V_p(w_{\text{CVaR}}, P) = \left(\sum_{i=1}^{N-1} |w_{\text{CVaR}}(x_{i+1}) - w_{\text{CVaR}}(x_i)|^p\right)^{\frac{1}{p}} = |w_{\text{CVaR}}(x_{i+1}) - w_{\text{CVaR}}(x_i)| = \alpha^{-1}.$$

Thus, we see $V_p(w_{\text{CVaR}}, P) \leq \alpha^{-1}$ for all partitions P, which proves $V_p(w_{\text{CVaR}}) = \alpha^{-1}$. □

Proof of Bounded Variation of Trimmed Risk. Let $P = \{x_1, \ldots, x_N\}$ be a partition such that $x_1 \leq \ldots \leq x_N$. Note that the scoring function of trimmed risk is $w_{\text{TRIM}}(t) = w(t) = \frac{1}{1-2\alpha}\mathbf{1}\{\alpha \leq t \leq 1-\alpha\}$ for $t \in [0,1]$. There are three cases.

Case 1: we have $P \subset [\alpha, 1-\alpha)$ or $P \subset [0,\alpha) \cup [1-\alpha, 1]$. In this case, $w_{\text{TRIM}}(x_i) = w_{\text{TRIM}}(x_j)$ for all $i, j = 1, \ldots, N$. Thus, we see $V_p(w_{\text{TRIM}}, P) = 0$.

Case 2: we have a) $x_1 \geq \alpha$ and an i for which $x_i \leq 1-\alpha$ and $x_{i+1} > 1-\alpha$; or b) $x_1 \leq 1-\alpha$ and an i for which $x_i < \alpha$ and $x_{i+1} \geq \alpha$. Here, we obtain

$$V_p(w_{\text{TRIM}}, P) = \left(\sum_{i=1}^{N-1} |w_{\text{TRIM}}(x_{i+1}) - w_{\text{TRIM}}(x_i)|^p\right)^{\frac{1}{p}} = |w_{\text{TRIM}}(x_{i+1}) - w_{\text{TRIM}}(x_i)| = (1-2\alpha)^{-1}.$$

Case 3: we have i, j such that $x_i < \alpha$, $x_{i+1} \geq \alpha$, $x_j \leq 1-\alpha$ and $x_j > 1-\alpha$. In this setting, we have

$$V_p(w_{\text{TRIM}}, P) = \left(\sum_{i=1}^{N-1} |w_{\text{TRIM}}(x_{i+1}) - w_{\text{TRIM}}(x_i)|^p\right)^{\frac{1}{p}} = (2|w_{\text{TRIM}}(x_{i+1}) - w_{\text{TRIM}}(x_i)|^p)^{\frac{1}{p}} = 2^{\frac{1}{p}}(1-2\alpha)^{-1}.$$

Thus, we see $V_p(w_{\text{TRIM}}, P) \leq 2^{\frac{1}{p}}(1-2\alpha)^{-1}$ for all partitions P, which proves $V_p(w_{\text{TRIM}}) = 2^{\frac{1}{p}}(1-2\alpha)^{-1}$. □

G. Additional Plots

In this section, we provide an additional plot of training curves to show that our optimization heuristic improves much faster than optimizing the precise dual form of CVaR. However, as noted previously, the precise dual form leads to a marginally superior solution.

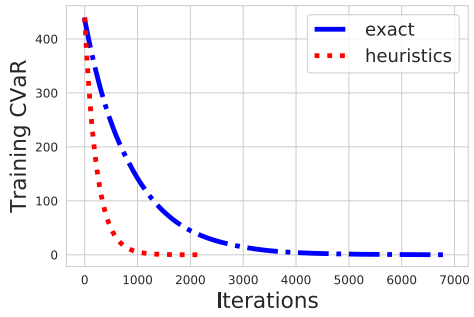


Figure 5. Convergence performance when optimizing CVaR in the linear regression setting using our heuristic method and an exact method that utilizes the convex dual form of CVaR in Equation (8). Both methods are used with the same fixed step size. Although our heuristic method improves faster, the exact method reaches a lower minimum.