Variance-based Regularization with Convex Objectives

Anonymous Author(s)
Affiliation
Address
Email

Abstract

We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error. Our approach builds off of techniques for distributionally robust optimization and Owen’s empirical likelihood, and we provide a number of finite-sample and asymptotic results characterizing the theoretical performance of the estimator. In particular, we show that our procedure comes with certificates of optimality, achieving (in some scenarios) faster rates of convergence than empirical risk minimization by virtue of automatically balancing bias and variance. We give corroborating empirical evidence showing that in practice, the estimator indeed trades between variance and absolute performance on a training sample, improving out-of-sample (test) performance over standard empirical risk minimization for a number of classification problems.

1 Introduction

Let \( \mathcal{X} \) be a sample space, \( P \) a distribution on \( \mathcal{X} \), and \( \Theta \) a parameter space. For a loss function \( \ell : \Theta \times \mathcal{X} \to \mathbb{R} \), consider the problem of finding \( \theta \in \Theta \) minimizing the risk

\[
R(\theta) := \mathbb{E}[\ell(\theta, X)] = \int \ell(\theta, x) dP(x)
\]

(1)

given a sample \( \{X_1, \ldots, X_n\} \) drawn i.i.d. according to the distribution \( P \). Under appropriate conditions on the loss \( \ell \), parameter space \( \Theta \), and random variables \( X \), a number of researchers \( \{1, 6, 14, 7, 3\} \) give results of the form

\[
R(\theta) \leq \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) + C_1 \sqrt{\frac{\text{Var}(\ell(\theta, X))}{n}} + C_2 \frac{1}{n}
\]

(2)

holds with high-probability, where \( C_1 \) and \( C_2 \) depend on the parameters of problem \( \{1\} \) and the desired confidence guarantee. Such bounds justify empirical risk minimization, which chooses \( \hat{\theta}_n \) to minimize \( \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) \) over \( \theta \in \Theta \). Further, these bounds showcase a tradeoff between bias and variance, where we identify the bias (or approximation error) with the empirical risk \( \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) \), while the variance arises from the second term in the bound.

Considering the bias-variance tradeoff \( \{1\} \) in statistical learning, it is natural to instead choose \( \theta \) to directly minimize a quantity trading between approximation and estimation error:

\[
\frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) + C \sqrt{\frac{\text{Var}_n(\ell(\theta, X))}{n}},
\]

(3)

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.
where $\text{Var}_{\hat{P}_n}$ denotes the empirical variance. Maurer and Pontil [19] consider this idea, giving a number of guarantees on the convergence and good performance of such a procedure. Unfortunately, even when the loss $\ell$ is convex in $\theta$, the formulation (3) is generally non-convex, yielding computationally intractable problems, which has limited the applicability of procedures that minimize the variance-corrected empirical risk (3). In this paper, we develop an approach based on Owen’s empirical likelihood [22] and ideas from distributionally robust optimization [4][5][12] that—whenever the loss $\ell$ is convex—provides a tractable convex formulation closely approximating the penalized risk (3), and we give a number of theoretical guarantees and empirical evidence for its performance.

To describe our approach, we require a few definitions. For a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ with $\phi(1) = 0$, $D_\phi(P||Q) = \int_X \phi\left(\frac{dP}{dQ}\right) dQ$ is the $\phi$-divergence between distributions $P$ and $Q$ defined on $X$. Throughout this paper, we use $\phi(t) = \frac{1}{2}(t - 1)^2$, which gives the $\chi^2$-divergence. Given $\phi$ and an i.i.d. sample $X_1, \ldots, X_n$, we define the $\rho$-neighborhood of the empirical distribution 

$$\mathcal{P}_n := \left\{ \text{distributions } P \text{ s.t. } D_\phi(P||\hat{P}_n) \leq \frac{\rho}{n} \right\},$$

where $\hat{P}_n$ denotes the empirical distribution of the sample $\{X_i\}_{i=1}^n$, and our choice $\phi(t) = \frac{1}{2}(t - 1)^2$ means that $\hat{P}_n$ has support $\{X_i\}_{i=1}^n$. We then define the robustly regularized risk

$$R_n(\theta, \mathcal{P}_n) := \sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(\theta, X)] = \sup_P \left\{ \mathbb{E}_P[\ell(\theta, X)] : D_\phi(P||\hat{P}_n) \leq \frac{\rho}{n} \right\}. \quad (4)$$

As it is the supremum of a family of convex functions, the robust risk $\theta \mapsto R_n(\theta, \mathcal{P}_n)$ is convex in $\theta$ regardless of the value of $\rho \geq 0$ whenever the original loss $\ell(\cdot; X)$ is convex and $\Theta$ is a convex set. Namkoong and Duchi [21] propose a stochastic procedure for minimizing (4) almost as fast as stochastic gradient descent. See Appendix C for a detailed account of an alternative method.

We show that the robust risk (4) provides an excellent surrogate for the variance-regularized quantity (3) in a number of ways. Our first result (Thm. 1 in Sec. 2) is that for bounded loss functions,

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2\rho}{n}} \text{Var}_{\hat{P}_n}(\ell(\theta, X)) + \varepsilon_n(\theta), \quad (5)$$

where $\varepsilon_n(\theta) \leq 0$ and $O(1/n)$ uniformly in $\theta$. We show that when $\ell(\theta, X)$ has suitably large variance, with high probability $\varepsilon_n = 0$. With the expansion (5) in hand, we can show a number of finite-sample convergence guarantees for the estimator

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \left\{ \sup_{P} \left\{ \mathbb{E}_P[\ell(\theta, X)] : D_\phi(P||\hat{P}_n) \leq \frac{\rho}{n} \right\} \right\}. \quad (6)$$

Based on the expansion (5), solutions $\hat{\theta}_n$ of problem (6) enjoy automatic finite sample optimality certificates: for $\rho \geq 0$, with probability at least $1 - C_1 \exp(-\rho)$ we have

$$\mathbb{E}[\ell(\hat{\theta}_n; X)] \leq R_n(\hat{\theta}_n; \mathcal{P}_n) + \frac{C_2\rho}{n} = \inf_{\theta \in \Theta} R_n(\theta; \mathcal{P}_n) + \frac{C_2\rho}{n},$$

where $C_1, C_2$ are constants (which we specify) that depend on the loss $\ell$ and domain $\Theta$. That is, with high probability the robust solution has risk no worse than the optimal finite sample robust objective up to an $O(\rho/n)$ error term. To guarantee a desired level of risk performance with probability $1 - \delta$, we may specify the robustness penalty $\rho = O(\log \frac{1}{\delta})$. Secondly, we show that the procedure (6) allows us to automatically and near-optimally trade between approximation and estimation error (bias and variance), so that

$$\mathbb{E}[\ell(\hat{\theta}_n; X)] \leq \inf_{\theta \in \Theta} \left\{ \mathbb{E}[\ell(\theta; X)] + 2 \sqrt{\frac{2\rho}{n}} \text{Var}(\ell(\theta; X)) \right\} + \frac{C\rho}{n},$$

with high probability. When there are parameters $\theta$ with small risk $R(\theta)$ (relative to the optimal parameter $\theta^*$) and small variance $\text{Var}(\ell(\theta, X))$, this guarantees that the excess risk $R(\hat{\theta}_n) - R(\theta^*)$ is essentially of order $O(\rho/n)$, where $\rho$ governs our desired confidence level. We give an explicit example in Section 3.2 where our robustly regularized procedure (6) converges at $O(\log n/n)$ compared to $O(1/\sqrt{n})$ of empirical risk minimization.
Bounds that trade between risk and variance are known in a number of cases in the empirical risk minimization literature [18, 25, 1, 71, 6, 7, 14], which is relevant when one wishes to achieve “fast rates” of convergence for statistical learning algorithms. In many cases, such tradeoffs require either conditions such as the Mammen-Tsybakov noise condition [18, 6] or localization results [3, 1, 20].

We also provide a uniform variant Theorem 1. This result, and our subsequent results, depend on where \( \epsilon \). These results validate our theoretical predictions, showing that the robust solutions are a practical convergence under typical curvature conditions on the risk \( R \).

We complement our theoretical results in Section 4 where we conclude by providing two experiments comparing empirical risk minimization (ERM) strategies to robustly-regularized risk minimization (\( R_\rho \)). These results validate our theoretical predictions, showing that the robust solutions are a practical alternative to empirical risk minimization. In particular, we observe that the robust solutions outperform their ERM counterparts on “harder” instances with higher variance. In classification problems, for example, the robustly regularized estimators exhibit an interesting tradeoff, where they improve performance on rare classes (where ERM usually sacrifices performance to improve the common cases—increasing variance slightly) at minor cost in performance on common classes.

2 Variance Expansion

We begin our study of the robust regularized empirical risk \( R_{\rho} (\theta, P_n) \) by showing that it is a good approximation to the empirical risk plus a variance term [1]. Although the variance of the loss is in general non-convex, the robust formulation [\( \rho \)] is a convex optimization problem for variance regularization whenever the loss function is convex [cf. 13, Prop. 2.1.2.]. First, we show a general result on the robust empirical risk, recalling that \( \phi (t) = \frac{1}{2} (t - 1)^2 \) in our definition of the \( \phi \)-divergence.

**Theorem 1.** Let \( Z \) be a random variable taking values in \([-M, M]\) and fix \( \rho \geq 0 \). Then

\[
\left( \frac{2 \rho}{n} \text{Var} \hat{\rho}_n (Z) - \frac{2 M \rho}{n} \right) + \leq \sup_{P : D_\phi (P | \hat{P}_n) \leq \rho} \left\{ \mathbb{E}_P [Z] : D_\phi (P | \hat{P}_n) \leq \frac{\rho}{n} \right\} \leq \mathbb{E} \hat{\rho}_n [Z] \leq \sqrt{\frac{2 \rho}{n} \text{Var} \hat{\rho}_n (Z)},
\]

where

\[
\frac{24 \rho}{\sqrt{n} \text{Var}(Z)} \leq \frac{16}{\sqrt{n}} M^2 \text{ and we set } t_n = \sqrt{\frac{\text{Var}(Z)}{1 - n^{-1} - \frac{1}{2}}} - M^2 \geq \sqrt{\frac{\text{Var}(Z)}{18},}
\]

If \( n \geq \max \{ \frac{24 \rho}{\sqrt{n} \text{Var}(Z)}, \frac{16}{\sqrt{n}}, 1 \} M^2 \) and we set \( t_n = \sqrt{\text{Var}(Z)} \left( \sqrt{1 - n^{-1} - \frac{1}{2}} - M^2 \right) \), then

\[
\mathbb{E} P [Z] = \mathbb{E} \hat{\rho}_n [Z] + \sqrt{\frac{2 \rho}{n} \text{Var} \hat{\rho}_n (Z)} \leq \mathbb{E} \hat{\rho}_n [Z] + \sqrt{\frac{2 \rho}{n} \text{Var} \hat{\rho}_n (Z)},
\]

with probability at least \( 1 - \exp \left( - \frac{n t_n^2}{2 M^2} \right) \geq 1 - \exp \left( - \frac{n \text{Var}(Z)}{36 M^2} \right) \).

See Appendix [A.1] for the proof.

Inequality (7) and the exact expansion (8) show that, at least for bounded loss functions \( \ell \), the robustly regularized risk (4) is a natural (and convex) surrogate for empirical risk plus standard deviation of the loss, and the robust formulation approximates exact variance regularization with a convex penalty.

We also provide a uniform variant Theorem 1. This result, and our subsequent results, depend on localized Rademacher averages. Recall that the empirical Rademacher complexity of a function class \( \mathcal{F} \subset \{ f : \mathcal{X} \to \mathbb{R} \} \),

\[
\mathcal{R}_n (\mathcal{F}) := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f (X_i)
\]

where \( \epsilon_i \in \{-1, 1\} \) are i.i.d. random signs. Although we state our results abstractly, we typically take \( \mathcal{F} := \{ \ell (\theta, \cdot) : \theta \in \Theta \} \). Our results depend on the localized Rademacher complexities of subsets of the function class \( \mathcal{F} \); such localized Rademacher averages typically give tighter error bounds than their global counterparts [1]. The following definition will be useful.

**Definition 2.1.** A function \( \psi : \mathbb{R} \to \mathbb{R}_+ \) is sub-root if it is nonnegative, nondecreasing and \( r \mapsto \psi (r) / \sqrt{r} \) is nonincreasing for all \( r > 0 \).

Letting \( \psi_n : \mathbb{R}_+ \to \mathbb{R}_+ \) satisfy \( \psi_n (r) \geq \mathbb{E} [\mathcal{R}_n (\{ c f : f \in \mathcal{F}, c \in [0, 1], \mathbb{E} [c^2 f^2] \leq r \})] \), it is often possible to choose \( \psi_n \) to be close to \( \mathbb{E} [\mathcal{R}_n] \) and sub-root. If \( \psi_n \) is a non-constant sub-root function,
then it is continuous and has a unique positive fixed point \( \psi_n(r^*_n) = r^*_n \), which plays a fundamental role in providing uniform convergence guarantees [1, Lemma 3.2]. For example, when \( \mathcal{F} \) is a VC-class, we typically have \( r^*_n \leq \frac{\text{VC}(\mathcal{F}) \log n}{n} \) (see Lemma 3.1 to come); more generally, whenever \( \mathcal{F} = \{ \ell(\theta, \cdot) \mid \theta \in \Theta \} \) for a compact set \( \Theta \subseteq \mathbb{R}^d \) with diameter \( \text{diam}(\Theta) \), chaining arguments [27] Chapter 2] typically give \( \mathbb{E}[\mathcal{R}_n(\mathcal{F})] \leq \sqrt{\frac{2 \text{diam}(\Theta)}{n}} \). In our theorems to follow, we can generally replace \( r^*_n \) with an upper bound of the form \( A/\sqrt{n} \) for a problem-dependent constant \( A \).

We are now ready to show that the variance expansion \(^3\) holds uniformly. For \( M \geq 1 \), let \( \mathcal{F} \) be a collection of bounded functions \( f : \mathcal{X} \to [-M, M] \) such that
\[
\mathbb{E} \left[ \mathcal{R}_n \left( \{ cf : f \in \mathcal{F}, c \in [0, 1], \mathbb{E}[e^f] \leq r \} \right) \right] \leq \psi_n(r) \tag{9}
\]
for a sub-root function \( \psi_n(\cdot) \) with \( r^*_n \geq \psi_n(r^*_n) \). Define the set \( \mathcal{F}_{\geq \tau} := \{ f \in \mathcal{F} : \text{Var}(f) \geq \tau^2 \} \).

**Theorem 2.** Let \( \tau > 0 \) and \( t > \log \left( \frac{4nM^2}{t} + 1 \right) \). Assume that \( n \) satisfies
\[
\frac{\tau n}{16} \geq M^2 \rho + 9 \left( 17M^2 \vee \tau \right) nr^*_n + (5 (M \vee 6\tau) M + (30\tau \vee 1)) t.
\]
Then, with probability at least \( 1 - e^{-t} \), for all \( f \in \mathcal{F}_{\geq \tau} \)
\[
\sup_{P : D_n(P|\hat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[f(X)] = \mathbb{E}_{\hat{P}_n}[f(X)] + \sqrt{\frac{2\rho}{n} \text{Var}_{\hat{P}_n}(f(X))} \tag{10}
\]
We prove the theorem in Section A.3. Theorem 2 shows that the variance expansion of Theorem 1 holds uniformly for all functions \( f \) with sufficient variance.

### 3 Optimization by Minimizing the Robust Loss

Based on the variance expansions in the preceding section, we show that the robust solution \(^6\) automatically trades between approximation and estimation error. Our main result gives a general guarantee, which we immediately apply to the risk minimization problem.

For \( M \geq 1 \), let \( \mathcal{F} \) be a collection of bounded functions \( f : \mathcal{X} \to [-M, M] \) satisfying \(^9\). Define the empirical minimizer \( \hat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \sup_P \left( \mathbb{E}_P[f(X)] : D_n(P|\hat{P}_n) \leq \frac{\rho}{n} \right) \right\} \).

**Theorem 3.** For an arbitrary fixed constant \( K > 2 \) and \( t \geq \log \left( \frac{4nM^2}{t} + 1 \right) \vee 1 \), let
\[
\frac{n}{8} \geq \frac{\rho}{6} \geq 16t (3M + 2t) + 9nr^*_n.
\]
Then, we have with probability at least \( 1 - 6e^{-t} \)
\[
\mathbb{E}[\hat{f}] \leq \sup_{P : D_n(P|\hat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[\hat{f}(X)] + 6r^*_n + 48M \sqrt{\frac{n}{n} \sqrt{r^*_n} + \frac{\sqrt{\rho}}{n} \left( 9M \sqrt{t} + 2\sqrt{\rho} \right)} \tag{11a}
\]
\[
\leq \inf_{f \in \mathcal{F}} \left\{ \mathbb{E}[f] + 6 \sqrt{\frac{\rho}{n} \text{Var}(f)} + 24\sqrt{\rho}r^*_n + 84M \sqrt{\frac{\rho}{n} \sqrt{r^*_n} + \frac{2\sqrt{\rho}}{n} \left( 11M \sqrt{t} + \sqrt{\rho} \right)} \right\}. \tag{11b}
\]
See Section A.2 for the proof. Unlike the analogous result [1] for empirical risk minimization, Theorem 3 does not require the self-bounding type assumption \( \text{Var}(f) \leq B \mathbb{E}[f] \). This is a consequence of using a peeling argument on the self-normalized class of functions
\[
\mathcal{G}_r = \left\{ r(\mathbb{E}[f^2] \vee r)^{-r}(f - \mathbb{E}[f]) : f \in \mathcal{F} \right\}
\]
as opposed to the variance-normalized class \( \mathcal{G}_r = \{ r(\mathbb{E}[f^2] \vee r)^{-1}f : f \in \mathcal{F} \} \) as in Bartlett et al. [1, Theorem 3.1]. Further, we note that the robust objective \( \sup_{P : D_n(P|\hat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[\hat{f}(X)] \) is an empirical likelihood upper confidence bound on the optimal population risk \([12]\). Since empirical likelihood confidence bounds are self-normalizing [22], there may exist a deeper connection whose investigation we leave to future work.
3.1 Consequences of Theorem 3

We now turn to a number of corollaries that expand on Theorem 3. Our first corollary shows that Theorem 3 applies to standard Vapnik-Chervonenkis (VC) classes. As VC dimension is preserved through composition, this result also extends to the procedure (5) in typical empirical risk minimization scenarios. We prove the below lemma in Section A.4.

**Lemma 3.1.** For $M > 1$, let $F$ be a collection of bounded functions $f : \mathcal{X} \to [-M, M]$ with VC-dimension VC$(F)$ and let $\psi_n(r) = \mathbb{E} R_n\left\{cf : f \in F, c \in [0, 1], E[c^2 f^2] \leq r \right\}$. Then, $\hat{\psi}_n(r)$ is sub-root and its fixed point is bounded by $r_n^* \leq CMVC(F) \log n / n$ for a numerical constant $C > 0$.

We now focus more explicitly on the estimator $\hat{\theta}_n$ defined by minimizing the robust regularized risk $\ell$. Let us assume that $\Theta \subseteq \mathbb{R}^d$, and that we have a typical linear modeling situation, where a loss $h$ is applied to an inner product, that is, $\ell(\theta, x) = h(\theta^\top x)$. In this case, by making the substitution that the class $F = \{\ell(\theta, \cdot) : \theta \in \Theta\}$ in Theorem 3, we have VC$(F) \leq d$, and we obtain the following corollary.

**Corollary 3.1.** Let the conditions of the previous paragraph hold and let $\hat{\hat{\theta}}_n \in \arg\min_{\theta \in \Theta} R_n(\theta, P_n)$. Assume also that $\ell(\theta, x) \in [0, M]$ for all $\theta \in \Theta, x \in \mathcal{X}$. Then, if $n \geq 2\rho \geq C_M d \log n$,

$$R(\hat{\hat{\theta}}_n) \leq R_n(\hat{\hat{\theta}}_n, P_n) + C_2 \rho / n \leq \inf_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{\rho}{n} \text{Var}(\ell(\theta, X))} \right\} + C_3 \rho n$$

with probability at least $1 - \frac{C_4}{n}$ where $C_i$’s are numerical constants.

Unpacking Theorem 3 and Corollary 3.1 a bit, the first result (11a) (and its counterpart (12)) provides a high-probability guarantee that the true expectation $\mathbb{E}\ell$ cannot be more than $O(1/n)$ worse than its robustly-regularized empirical counterpart, that is, $R(\hat{\theta}_n) \leq R_n(\hat{\theta}_n, P_n) + O(\rho/n)$, which is (roughly) a consequence of uniform variant of Bernstein’s inequality. The second result (11b) (and inequality (12) from Corollary 3.1) guarantee the convergence of the empirical minimizer to a parameter with risk at most $O(1/n)$ larger than the best possible variance-corrected risk. In the case that the losses take values in $[0, M]$, then $\text{Var}(\ell(\theta, X)) \leq MR(\theta)$, and Theorem 3 implies

$$R(\hat{\theta}_n) \leq R(\theta^*) + C \sqrt{\frac{M M R(\theta^*) n}{n}} + C_3 \rho n$$

a type of result achieved by empirical risk minimization for bounded nonnegative losses [6, 29, 23]. In some scenarios, however, the variance may satisfy $\text{Var}(\ell(\theta, X)) \ll MR(\theta)$, yielding improvements.

To give an alternative variant of Corollary 3.1, let $\Theta \subseteq \mathbb{R}^d$ and assume that for each $x \in \mathcal{X}$, $\ell$ is $L$-Lipschitz in $\theta$. If $D := \text{diam}(\Theta) = \sup_{\theta, \theta' \in \Theta} ||\theta - \theta'|| < \infty$, then $\ell(\theta, x) \leq L \text{diam}(\Theta) = : M$. As in Lemma 3.1, we have the below bound on the fixed point $r_n^*$. See Section A.5 for the proof.

**Lemma 3.2.** Let the conditions of the previous paragraph hold and let $\hat{\psi}_n(r) = \mathbb{E} R_n\left\{cf : f \in F, c \in [0, 1], E[c^2 f^2] \leq r \right\}$. Then, $\hat{\psi}_n(r)$ is sub-root and its fixed point is bounded by $r_n^* \leq CL \text{diam}(\Theta) \sqrt{\frac{n}{n}}$ for a numerical constant $0 < C < \infty$.

We conclude that the same result as in Corollary 3.1 holds with $M = L \text{diam}(\Theta)$ in the bound (12).

3.2 Beating empirical risk minimization

We now provide an example in which the robustly-regularized estimator (5) exhibits a substantial performance gap over empirical risk minimization. We expect the robust approach to offer performance benefits in situations in which the empirical risk minimizer is highly sensitive to noise, say, because the losses are piecewise linear, and slight under- or over-estimates of slope may significantly degrade solution quality. With this in mind, we construct a toy 1-dimensional example—estimating the median of a distribution supported on $\mathcal{X} = \{-1, 0, 1\}$—in which the robust-regularized estimator has convergence rate $\log n / n$, while empirical risk minimization is at best $1/\sqrt{n}$.

Define the loss $\ell(\theta, x) = |\theta - x| - x$, and for $\delta \in (0, 1)$ let the distribution $P$ be defined by $P(X = 1) = \frac{1-\delta}{2}$, $P(X = -1) = \frac{1+\delta}{2}$, $P(X = 0) = \delta$. Then for $\theta \in \mathbb{R}$, the risk of the loss is $R(\theta) = \delta|\theta| + \frac{1-\delta}{2} |\theta-1| + \frac{1-\delta}{2} |\theta+1| - (1-\delta)$. By symmetry, it is clear that $\theta^* := \arg\min_\theta R(\theta) = 0$,
which satisfies $R(\theta^*) = 0$. (Note that $\ell(\theta, x) = \ell(\theta, x) - \ell(\theta^*, x)$.) Without loss of generality, we assume that $\Theta = [-1,1]$. Define the empirical risk minimizer and the robust solution

$$\hat{\theta}_{\text{erm}} := \arg\min_{\theta \in \mathbb{R}} \mathbb{E}_{\rho_n} [\ell(\theta, X)] = \arg\min_{\theta \in [-1,1]} \mathbb{E}_{\rho_n} [||\theta - X||], \quad \hat{\theta}_{\text{rob}} \in \arg\min_{\theta \in \Theta} R_n(\theta, P_n).$$

Intuitively, if too many of the observations satisfy $X_i = 1$ or too many satisfy $X_i = -1$, then $\hat{\theta}_{\text{erm}}$ will be either 1 or −1; for small $\delta$, such events become reasonably probable. On the other hand, we have $\ell(\theta^*; x) = 0$ for all $x \in X$, so that $\mathbb{V}(\ell(\theta^*; X)) = 0$ and variance regularization achieves the rate $O(\log(n)/n)$ as opposed to empirical risk minimizer’s $O(1/\sqrt{n})$. See Section A.3 for the proof.

**Proposition 1.** Under the conditions of the previous paragraph, for $n \geq \rho \geq C_1 \log n$, with probability at least $1 - C_2 / n$, we have $R(\hat{\theta}_{\text{rob}}) - R(\theta^*) \leq C_3 / n$. However, with probability at least

$$2\Phi(-\sqrt{\frac{n}{n-1}}) - 2\sqrt{2/\pi n} \geq 2\Phi(-\sqrt{\frac{n}{n-1}}) - n^{-\frac{1}{2}},$$

we have $R(\hat{\theta}_{\text{erm}}) \geq R(\theta^*) + n^{-\frac{1}{2}}$.

For $n \geq 20$, the probability of the latter event is $\geq 0.088$. Hence, for this (specially constructed) example, we see that there is a gap of nearly $n^{\frac{1}{2}}$ in order of convergence.

### 3.3 Fast Rates

In cases in which the risk $R$ has curvature, empirical risk minimization often enjoys faster rates of convergence [6, 24]. The robust solution $\hat{\theta}_{\text{rob}}$ similarly attains faster rates of convergence in such cases, even with approximate minimizers of $R_n(\theta, P_n)$. For the risk $R$ and $\epsilon \geq 0$, let

$S^c := \{ \theta \in \Theta : R(\theta) \leq \inf_{\theta' \in \Theta} R(\theta') + \epsilon \}$

denote the $\epsilon$-sub-optimal (solution) set, and similarly let $S'_c := \{ \theta \in \Theta : R_n(\theta, P_n) \leq \inf_{\theta' \in \Theta} R_n(\theta', P_n) + \epsilon \}$. For a vector $\theta \in \Theta$, let $\pi_A(\theta) = \arg\min_{\theta \in S} ||\theta - \theta||_2$ denote the Euclidean projection of $\theta$ onto the set $S$, and for $A \subset \Theta$ let $\mathcal{R}_n(A)$ denote the Rademacher complexity of the localized process $\{x \mapsto \ell(\theta; x) - \ell(\pi_A(\theta); x) : \theta \in A\}$.

We then have the following result, whose proof we provide in Section A.7.

**Theorem 4.** Let $\Theta \subset \mathbb{R}^d$ be convex and let $\ell(\cdot; x)$ be convex and $L$-Lipschitz for all $x \in X$. For constants $\lambda > 0$, $\gamma > 1$, and $r > 0$, assume that $R$ satisfies

$$R(\theta) - \inf_{\theta \in \Theta} R(\theta) \geq \lambda \text{dist}(\theta, S^c) \gamma \text{ for all } \theta \text{ such that } \text{dist}(\theta, S) \leq r.$$  \hspace{1cm} (13)

Let $t > 0$. If $0 \leq \epsilon \leq \frac{1}{\lambda}$. Then

$$\epsilon \geq \left( \frac{8L^2}{n} \right)^{\frac{1}{\gamma + 1}} \left( \frac{2}{\lambda} \right) \frac{1}{\gamma} \text{ and } \frac{\epsilon}{2} \geq \frac{2}{\sqrt{\lambda}^n} \mathbb{E}[\mathcal{R}_n(S^c)] + L \left( \frac{2e}{\lambda} \right)^{\frac{1}{\gamma}} \sqrt{\frac{2\pi}{n}},$$

then $\mathbb{P}(S^c \subset S^c) \geq 1 - e^{-t}$, and inequality (14) holds for all $\epsilon \geq \left( \frac{L^2(1 + d/d)}{\lambda \sqrt{\gamma}} \right)^{\frac{1}{\gamma + 1}}$.

### 4 Experiments

We present two real classification experiments to carefully compare standard empirical risk minimization (ERM) to the variance-regularized approach we present. We might expect that the ERM estimator $\hat{\theta}_{\text{erm}}$ performs poorly on rare classes with (relatively) more variance, which the robust solution reduces by via improved classification performance on rare instances. In our experiments, this indeed occurs, and with little expense over the more common instances.

#### 4.1 Protease cleavage experiments

For our first experiment, we compare our robust regularization procedure to other regularizers using the HIV-1 protease cleavage dataset from the UCI ML-repository [17]. In this binary classification task, one is given a string of amino acids (a protein) and a featurized representation of the string of dimension $d = 50960$. And the goal is to predict whether the HIV-1 virus will cleave the amino acid sequence in its central position. We have a sample of $n = 6590$ observations of this process, where the class labels are somewhat skewed: there are 1300 examples with label $Y = +1$ (HIV-1 cleaves) and 5230 examples with $Y = -1$ (does not cleave).
We use the logistic loss \( \ell(\theta; (x, y)) = \log(1 + \exp(-y\theta^T x)) \). We compare the performance of different constraint sets \( \Theta \) by taking \( \Theta = \{ \theta \in \mathbb{R}^d : a_1 \| \theta \|_1 + a_2 \| \theta \|_2 \leq r \} \), which is equivalent to elastic net regularization [30], while varying \( a_1, a_2, \) and \( r \). We experiment with \( \ell_1\)-constrained \((a_1 = 1, a_2 = 0)\) with \( r \in \{50, 100, 500, 1000, 5000\} \), \( \ell_2\)-constrained \((a_1 = 0, a_2 = 1)\) with \( r \in \{5, 10, 50, 100, 500\} \), elastic net \((a_1 = 1, a_2 = 10)\) with \( r \in \{10^2, 2 \cdot 10^2, 3 \cdot 10^2, 10^3, 10^4\} \), our robust regularizer with \( \rho \in \{10^2, 10^3, 10^4, 5 \cdot 10^4, 10^5\} \) and our robust regularizer coupled with the \( \ell_1\)-constraint \((a_1 = 1, a_2 = 0)\) with \( r = 100 \). Though we use a convex surrogate (logistic loss), we measure performance of the classifiers using the zero-one (misclassification) loss \( 1 \{ \text{sign}(\theta^T x) y \neq 0 \} \).

For validation, we perform 50 experiments, where in each experiment we randomly select 9/10 of the data to train the model, evaluating its performance on the held out 1/10 fraction (test).

We plot results summarizing these experiments in Figure 1. The horizontal axis in each figure indexes our choice of regularization value (so “Regularizer = 1” for the \( \ell_1\)-constrained problem corresponds to \( r = 50 \)). The figures show that the robustly regularized risk provides a different type of protection against overfitting than standard regularization or constraint techniques do: while other regularizers underperform in heavily constrained settings, the robustly regularized estimator \( \hat{\theta}^{\text{rob}} \) achieves low classification error for all values of \( \rho \). Notably, even when coupled with a fairly stringent \( \ell_1\)-constraint \((r = 100)\), robust regularization has performance better than \( \ell_1 \) except for large values \( r \), especially on the rare label \( Y = +1 \).

We investigate the effects of the robust regularizer with a slightly different perspective in Table 1 where we use \( \Theta = \{ \theta : \| \theta \|_1 \leq 100 \} \) for the constraint set for each experiment. We give error rates and logistic risk values for the different procedures, averaged over 50 independent runs. We note that all gaps are significant at the 3-standard error level. We see that the ERM solutions achieve good performance on the common class \((Y = -1)\) but sacrifice performance on the uncommon class. As we increase \( \rho \), performance of the robust solution \( \hat{\theta}^{\text{rob}} \) on the rarer label \( Y = +1 \) improves, while the error rate on the common class degrades a small (insignificant) amount.

| \( \rho \) | train risk | test risk | error \((Y = +1)\) | error \((Y = -1)\) |
|---|---|---|---|---|
| erm | 0.1587 | 0.1706 | 5.52 | 6.39 | 17.32 | 18.79 | 2.45 | 3.17 |
| 100 | 0.1623 | 0.1763 | 4.99 | 5.92 | 15.01 | 17.04 | 2.38 | 3.02 |
| 1000 | 0.1777 | 0.1944 | 4.5 | 5.92 | 13.35 | 16.33 | 2.19 | 3.2 |
| 10000 | 0.283 | 0.3031 | 2.39 | 5.67 | 7.18 | 14.65 | 1.15 | 3.32 |

4.2 Document classification in the Reuters corpus

For our second experiment, we consider a multi-label classification problem with a reasonably large dataset. The Reuters RCV1 Corpus [16] has 804,414 examples with \( d = 47,236 \) features, where feature \( j \) is an indicator variable for whether word \( j \) appears in a given document. The goal is to classify documents as a subset of the 4 categories where documents are labeled with a subset of those. As documents can belong to multiple categories, we fit binary classifiers on each of the four
categories. Each category has different number of documents (Corporate: 381, 327, Economics: 119, 920, Government: 239, 267, Markets: 204, 820). In this experiment, we expect the robust solution to outperform ERM on the rarer category (Economics), as the robustification \( \theta \) naturally upweights rarer (harder) instances, which disproportionally affect variance—as in the previous experiment.

For each category \( k \in \{1, 2, 3, 4\} \), we use the logistic loss \( \ell(\theta_k; (x, y)) = \log(1 + \exp(-y\theta_k^T x)) \). For each binary classifier, we use the \( \ell_1 \) constraint set \( \Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\} \). To evaluate performance on this multi-label problem, we use precision (ratio of the number of correct positive labels to the number classified as positive) and recall (ratio of the number of correct positive labels to the number of actual positive labels). We partition the data into ten equally-sized sub-samples and perform ten validation experiments, where in each experiment we use one of the ten subsets for fitting the logistic models and the remaining nine partitions as a test set to evaluate performance.

In Figure 2, we summarize the results of our experiment averaged over the 10 runs, with 2-standard error bars (computed across the folds). To facilitate comparison across the document categories, we give exact values of these averages in Tables 2 and 3. Both \( \hat{\theta}^{\text{rob}} \) and \( \hat{\theta}^{\text{erm}} \) have reasonably high precision across all categories, with increasing \( \rho \) giving a mild improvement in precision (from \( .93 \pm .005 \) to \( .94 \pm .005 \)). On the other hand, we observe in Figure 2(c) that ERM has low recall (.69 on test) for the Economics category, which contains about 15% of documents. As we increase \( \rho \) from 0 (ERM) to \( 10^3 \), we see a smooth and substantial improvement in recall for this rarer category (without significant degradation in precision). This improvement in recall amounts to reducing variance in predictions on the rare class. This precision and recall improvement comes in spite of the increase in the average binary logistic risk for each of the 4 classes. In Figure 2(a), we plot the average binary logistic risk (on train and test sets) averaged over the 4 categories as well as the upper confidence bound \( R_n(\theta, P_n) \) as we vary \( \rho \). The robust regularization effects reducing variance appear to improve the performance of the binary logistic loss as a surrogate for true error rate.

### Table 2: Reuters Corpus Precision (%)

| \( \rho \) | Precision | Corporate | Economics | Government | Markets |
|----------|-----------|-----------|-----------|------------|---------|
|         | train     | test      | train     | test       | train   | test    |
| erm     | 92.72     | 92.7      | 93.55     | 93.55      | 89.02   | 89      | 94.1     | 94.12    | 92.88   | 92.94   |
| 1E3     | 92.97     | 92.95     | 93.31     | 93.33      | 87.84   | 87.81   | 93.73    | 93.76    | 92.56   | 92.62   |
| 1E4     | 93.45     | 93.45     | 93.58     | 93.61      | 87.6    | 87.58   | 93.77    | 93.8     | 92.71   | 92.75   |
| 1E5     | 94.17     | 94.16     | 94.18     | 94.19      | 86.55   | 86.56   | 94.07    | 94.09    | 93.16   | 93.24   |
| 1E6     | 91.2      | 91.19     | 92        | 92.02      | 74.81   | 74.8    | 91.19    | 91.25    | 89.98   | 90.18   |

### Table 3: Reuters Corpus Recall (%)

| \( \rho \) | Recall | Corporate | Economics | Government | Markets |
|----------|--------|-----------|-----------|------------|---------|
|         | train  | test      | train     | test       | train   | test    |
| erm     | 90.97  | 90.96     | 90.20     | 90.25      | 67.53   | 67.56   | 90.49    | 90.49    | 88.77   | 88.78   |
| 1E3     | 91.72  | 91.69     | 90.83     | 90.86      | 70.42   | 70.39   | 91.26    | 91.23    | 89.62   | 89.58   |
| 1E4     | 92.40  | 92.39     | 91.47     | 91.54      | 72.38   | 72.36   | 91.76    | 91.76    | 90.48   | 90.45   |
| 1E5     | 93.46  | 93.44     | 92.65     | 92.71      | 76.79   | 76.78   | 92.26    | 92.21    | 91.46   | 91.47   |
| 1E6     | 93.10  | 93.08     | 92.00     | 92.04      | 79.84   | 79.71   | 91.89    | 91.90    | 92.00   | 91.97   |
References

[1] P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] P. L. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[4] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[5] D. Bertsimas, V. Gupta, and N. Kallus. Robust SAA. *arXiv:1408.4445 [math.OC]*, 2014. URL http://arxiv.org/abs/1408.4445.

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[8] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[9] O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, pages 213–247. Springer, 2003.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the $\ell_1$-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[12] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425 [stat.ML]*, 2016. URL http://arxiv.org/abs/1610.03425.

[13] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[14] V. Kolchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2636, 2006.

[15] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[16] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 27:1808–1829, 2001.

[17] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[18] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

[19] A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

[20] S. Mendelson. Learning without concentration. In *Proceedings of the Twenty Seventh Annual Conference on Computational Learning Theory*, 2014.

[21] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

[22] A. B. Owen. *Empirical likelihood*. CRC press, 2001.

[23] P. Samson. Concentration of measure inequalities for Markov chains and $\phi$-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

[24] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[25] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

[26] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[27] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[28] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[29] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

[30] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

[31] A. Zubkov and A. Serov. A complete proof of universal inequalities for the distribution function of the binomial law. *Theory of Probability & Its Applications*, 57(3):539–544, 2013.
A Proofs of Main Results

In this section, we provide the proofs of all of our major results. Within each proof, we defer arguments for more technical and ancillary results to appendices as necessary.

A.1 Proof of Theorem I

Let \( \sigma^2 = \text{Var}(Z) \) and \( s_n^2 = \text{Var}_{\hat{P}_n}(Z) = \mathbb{E}_{\hat{P}_n}[Z^2] - \mathbb{E}_{\hat{P}_n}[Z]^2 \) denote the population and sample variance of \( Z \), respectively.

Proof Overview

Consider the quadratically constrained linear maximization problem of interest

\[
\max_{\rho} \sum_{i=1}^{n} p_i z_i \quad \text{subject to} \quad p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}^n_{+} : \frac{1}{2} \| np - 1 \|_2^2 \leq \rho, \langle 1, p \rangle = 1 \right\}, \tag{15}
\]

where \( z \in \mathbb{R}^n \) is a vector. For simplicity, let \( s_n^2 = \frac{1}{n} \| z \|_2^2 - \langle \bar{z} \rangle^2 = \frac{1}{n} \| z - \bar{z} \|_2^2 \) denote the empirical “variance” of the vector \( z \), where \( \bar{z} = \frac{1}{n} \langle 1, z \rangle \) is the mean value of \( z \). Then by introducing the variable \( u = p - \frac{1}{n} 1 \), the objective in problem (15) satisfies \( \langle p, z \rangle = \bar{z} + \langle u, z \rangle = \bar{z} + \langle u, z - \bar{z} \rangle \) because \( \langle u, 1 \rangle = 0 \). Thus problem (15) is equivalent to solving

\[
\max_{u \in \mathbb{R}^n} \bar{z} + \langle u, z - \bar{z} \rangle \quad \text{subject to} \quad \| u \|_2^2 \leq \frac{2 \rho}{n^2}, \quad \langle 1, u \rangle = 0, \quad u \geq -\frac{1}{n}.
\]

Notably, by the Cauchy-Schwarz inequality, we have \( \langle u, z - \bar{z} \rangle \leq \sqrt{\frac{2 \rho}{n}} \| z - \bar{z} \|_2 / n = \sqrt{\frac{2 \rho s_n^2}{n}} \), and equality is attained if and only if

\[
u_i = \frac{\sqrt{\frac{2 \rho}{n}} (z_i - \bar{z})}{n \| z - \bar{z} \|_2} = \frac{\sqrt{\frac{2 \rho}{n}} (z_i - \bar{z})}{\sqrt{ns_n^2}}.
\]

Of course, it is possible to choose such \( u_i \) while satisfying the constraint \( u_i \geq -1/n \) if and only if

\[
\min_{i \in [n]} \frac{\sqrt{\frac{2 \rho}{n}} (z_i - \bar{z})}{\sqrt{ns_n^2}} \geq -1. \tag{16}
\]

Thus, if inequality (16) holds for the vector \( z \)—that is, there is enough variance in \( z \)—we have

\[
\sup_{p \in \mathcal{P}_n} \langle p, z \rangle = \bar{z} + \sqrt{\frac{2 \rho s_n^2}{n}}.
\]

For losses \( \ell(\theta, X) \) with enough variance relative to \( \ell(\theta, X_i) = \mathbb{E}_{\hat{P}_n}[\ell(\theta, X_i)] \), that is, those satisfying inequality (16), then, we have

\[
R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2 \rho \text{Var}_{\hat{P}_n}(\ell(\theta, X))}{n}}.
\]

A slight elaboration of this argument, coupled with the application of a few concentration inequalities, yields the next theorem. The theorem as stated applies only to bounded random variables, but in subsequent sections we relax this assumption by applying the characterization (16) of the exact expansion.

Proof

The theorem is immediate if \( s_n = 0 \) or \( \sigma^2 = 0 \), as in this case \( \sup_{P : D_{\text{KL}}(P \| \hat{P}_n) \leq \rho/n} \mathbb{E}_P[Z] = \mathbb{E}_{\hat{P}_n}[Z] = \mathbb{E}[Z] \). In what follows, we will thus assume that \( \sigma^2, s_n^2 > 0 \). Recall the maximization problem (15), which is

\[
\max_{p} \sum_{i=1}^{n} p_i z_i \quad \text{subject to} \quad p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}^n_{+} : \frac{1}{2} \| np - 1 \|_2^2 \leq \rho, \langle 1, p \rangle = 1 \right\},
\]
and the solution criterion \(16\), which guarantees that the maximizing value of problem \(15\) is 
\[
\bar{z} + \sqrt{2 \rho s_n^2 / n} \quad \text{whenever} \quad \sqrt{2 \rho z_i - \bar{z}} / \sqrt{n s_n^2} \geq -1.
\]
Letting \(z = Z\), then under the conditions of the theorem, we have \(|z_i - \bar{z}| \leq M\), and to satisfy inequality \(16\) it is certainly sufficient that
\[
2 \rho M^2 / n s_n^2 \leq 1, \quad \text{or} \quad n \geq 2 \rho M^2 / s_n^2, \quad \text{or} \quad s_n^2 \geq 2 \rho M^2 / n.
\]
Conversely, suppose that \(s_n^2 < 2 \rho M^2 / n\). Then we have \(2 \rho s_n^2 / n < 4 \rho^2 M^2 / n^2\), which in turn implies that
\[
\sup_{p \in P_n} \langle p, z \rangle \geq \frac{1}{n} (1, z) + \left( \sqrt{\frac{2 \rho s_n^2}{n} - \frac{2 \rho M^2}{n}} \right).
\]
Combining this inequality with the condition \(17\) for the exact expansion to hold yields the two-sided variance bounds \(7\).

We now turn to showing the high-probability exact expansion \(8\), which occurs whenever the sample variance is large enough by expression \(17\). To that end, we show that \(s_n^2\) is bounded from below with high probability. Define the event
\[
\mathcal{E}_n := \left\{ s_n^2 \geq \frac{1}{4} \sigma^2 \right\},
\]
and let \(n \geq \max \left\{ \frac{16 \rho}{\sigma^2}, \frac{16}{\sigma^2}, 1 \right\} M^2\). Then, we have on event \(\mathcal{E}_n\)
\[
n \geq \frac{16 \rho M^2}{\sigma^2} \geq \frac{16 \rho M^2}{4 s_n^2} = 4 \rho M^2 / s_n^2,
\]
so that the sufficient condition \(17\) holds and expression \(8\) follows. We now argue that the event \(\mathcal{E}_n\) has high probability via the following lemma, which is essentially an application of known concentration inequalities for convex functions coupled with a few careful estimates of the expectation of standard deviations.

**Lemma A.1.** Let \(Z_i\) be independent random variables taking values in \([M_0, M_1]\) with \(M = M_1 - M_0\), and let \(s_n^2 = \frac{1}{n} \sum_{i=1}^n Z_i^2 - \left( \frac{1}{n} \sum_{i=1}^n Z_i \right)^2\). For all \(t \geq 0\), we have
\[
\mathbb{P} \left( s_n \geq \sqrt{\mathbb{E}s_n^2} + t \right) \vee \mathbb{P} \left( s_n \leq \sqrt{\mathbb{E}s_n^2} - \frac{M^2}{n} - t \right) \leq \exp \left( -\frac{n t^2}{2 M^2} \right).
\]
See Section \(B.2\) for a proof of the lemma. When the \(Z_i\) are i.i.d., we obtain
\[
\mathbb{P} \left( s_n \geq \sigma \sqrt{1 - n^{-1} + t} \right) \vee \mathbb{P} \left( s_n \leq \sigma \sqrt{1 - n^{-1} - \frac{M^2}{n} - t} \right) \leq \exp \left( -\frac{n t^2}{2 M^2} \right)
\]
where \(\sigma^2 = \text{Var}(Z)\).

Now, substitute \(t = \sigma (\sqrt{1 - n^{-1}} - \frac{1}{2}) - \frac{M^2}{n}\) so that
\[
\sigma (1 - n^{-\frac{1}{2}}) - \frac{M^2}{n} - t = \frac{1}{2} \sigma.
\]
Note that \(\frac{M^2}{n} \leq \frac{M}{\sqrt{n}} \leq \sigma/4\) and since \(\sqrt{1 - n^{-1}} \geq 1 - \frac{1}{2n \sqrt{1 - n^{-1}}} \) and \(\sigma^2 \leq M^2 / 4\) by standard variance bounds, our choice of \(n\) also satisfies \(n \geq 16 M^2 / \sigma^2 \geq 64\). We thus have \(t / \sigma \geq 1 - \frac{1}{16 \sqrt{64} - \frac{1}{2} - \frac{1}{4} > \frac{1}{16}}\). We obtain
\[
\mathbb{P} (\mathcal{E}_n) \geq 1 - \exp \left( -\frac{n \sigma^2 ((1 - n^{-1})^{1/2} - 1/2 - M^2 / \sigma n)^2}{2 M^2} \right) \geq 1 - \exp \left( -\frac{n \sigma^2}{36 M^2} \right).
\]
This gives the result \(8\).
A.2 Proof of Theorem

We make repeated use of the following version of Talagrand’s concentration inequality for supremum of empirical processes due to Bousquet [9].

**Lemma A.2.** Let \( r > 0 \) and \( F \) be a class of functions that map \( X \) into \([a, b]\) such that for every \( f \in F \), \( \text{Var}(f(X)) \leq r \). Then, with probability at least \( 1 - e^{-t} \),

\[
\sup_{f \in F} \{E[f] - E\hat{p}_n[f]\} \leq \inf_{\alpha > 0} \left\{ 2(1 + \alpha)E[\mathcal{R}_nF] + \sqrt{\frac{2rt}{n}} + (b - a) \left( \frac{1}{3} + \frac{1}{\alpha} \right) \right\}
\]

and with probability at least \( 1 - 2e^{-t} \),

\[
\sup_{f \in F} \{E[f] - E\hat{p}_n[f]\} \leq \inf_{\alpha > 0} \left\{ \frac{2}{1 - \alpha} E[\mathcal{R}_nF] + \sqrt{\frac{2rt}{n}} + (b - a) \left( \frac{1}{3} + \frac{1}{\alpha} + \frac{1 + \alpha}{2\alpha(1 - \alpha)} \right) \right\}.
\]

The same statements hold with \( \sup_{f \in F} (E\hat{p}_n[f] - E[f]) \) replacing the left-hand side of the inequalities.

Furthermore, identical results hold for \( F_{\text{centered}} = \{f - E[f] : f \in F\} \) with \( E[\mathcal{R}_nF] \) instead of \( E[\mathcal{R}_nF_{\text{centered}}] \).

See Bartlett et al. [1, Thm 2.1] for a derivation of the above result from Bousquet [8, 9].

We first show the following version of uniform Bernstein’s inequality with Rademacher complexities.

We make repeated use of the following version of Talagrand’s concentration inequality for supremum

\[
\text{Var}(f(X)) \leq \text{Var}( \hat{p}_n(f) ) = \frac{1}{n} \sum_{t=1}^{n} \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left( \psi_i f(X_i) > t \right) \right)
\]

Next, we show an important extension of Lemma A.3 that replaces the Rademacher complexity term \( E[\mathcal{R}_nF] \) by a local quantity \( r_n^* \), the fixed point of \( \psi_n(r) \). To this end, we use another peeling argument and apply Lemma A.3 to the self-normalized class

\[
G_r := \left\{ \sqrt{\frac{r}{E[f^2]}} f : f \in F \right\} \subseteq \left\{ cf : f \in F, E[c^2 f^2] \leq r, c \in [0, 1] \right\}
\]

to obtain the following result.

**Lemma A.4.** Let \( F \) be a collection of bounded functions \( f : X \to [-M, M] \) satisfying \( \psi_n(r) \) for some \( \psi_n(\cdot) \). Then, with probability at least \( 1 - e^{-t} \), for every \( f \in F \),

\[
E[f] \leq E\hat{p}_n[f] + \sqrt{\frac{2\text{Var}(f)}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)}
\]

\[
+ c_1 \sqrt{\text{Var}(f) \left( r_n^* + \frac{8c_2M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right)}
\]

\[
+ c_1 \left( r_n^* + \frac{8c_2M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right). \quad (18)
\]

The same statement holds with the roles of \( E[f] \) and \( E\hat{p}_n[f] \) reversed. Furthermore, the same result holds for \( F_{\text{centered}} = \{f - E[f] : f \in F\} \) with the same constants and \( r_n^* \).

See Section B.2.2 for the proof.

Next, we give a lemma that upper bounds the \( \text{Var}(f) \) in Lemma A.4 by the sample variance

\[
\text{Var}\hat{p}_n(f) := E\hat{p}_n[f^2] - (E\hat{p}_n[f])^2.
\]
**Lemma A.5.** Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [-M, M]$ satisfying (9) for some $\psi_n(\cdot)$. Let $c_1 > \frac{1}{2\pi^2}$ and $K > 1$. Then, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$

$$\mathbb{E}[f^2] \leq \frac{K}{K-1} \mathbb{E}_{\hat{p}_n}[f^2] + 8c_2^2KM^2r_n^* + \frac{M^2t}{n}(2c_2 + K).$$

Also, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$

$$\mathbb{E}_{\hat{p}_n}[f^2] \leq \frac{K + 1}{K} \mathbb{E}[f^2] + 8c_2^2KM^2r_n^* + \frac{M^2t}{n}(2c_2 + K).$$

See Section 4.2.4 for the proof.

Now, let $K > 2$ be an arbitrary fixed constant and assume that

$$\frac{n}{4} \geq 16 \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \left( c_1c_2M + t + \log \left( \frac{n}{t} + 1 \right) \right) + 2c_2^2nr_n^*.$$

Using Lemmas A.4 and A.5, we claim that if

$$\rho \geq \frac{2K}{K-2} \left( c_1^2r_n^* + 8 \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \left( c_1c_2M + t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right),$$

then with probability at least $1 - e^{-3t}$, for every $f \in \mathcal{F}$

$$\mathbb{E}[f] \leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{\rho}{2n}} \text{Var}_{\hat{p}_n}(f) + c_1 \left( 1 + 2\sqrt{\frac{\rho}{n}} \right) r_n^* + 8c_1M \sqrt{M^2K\rho} \frac{K}{n}$$

$$\left. \frac{1}{c_1n} \right) 3M\sqrt{(2c_2 + K)t} + \frac{1}{c_1} \left( \sqrt{\rho} + \frac{2\rho}{n} \right).$$

(19)

In order to prove (19), we proceed by upper bounding the variance term in (18) by its empirical counterpart. Define

$$\kappa_1 := \sqrt{\frac{2e}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + c_1 \sqrt{r_n^* + \frac{8c_2M}{c_1n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)}}$$

$$\kappa_2 := c_1 \left( r_n^* + \frac{8c_2M}{c_1n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right)$$

for ease of notation.

**Lemma A.6.** Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [-M, M]$ satisfying (9) for some $\psi_n(\cdot)$. Let $c_1 > \frac{1}{2\pi^2}$ and $K > 1$. Then, with probability at least $1 - e^{-2t}$, for every $f \in \mathcal{F}$

$$\text{Var}(f) \leq \frac{1}{1 - 2\kappa_2^2 - 1/K} \text{Var}_{\hat{p}_n}(f)$$

$$+ \frac{K - 1}{1 - 2\kappa_2^2 - 1/K} \left( 32c_2^2KM^2r_n^* + \frac{4M^2t}{n}(2c_2 + K) \right) + \frac{2c_2^2}{1 - 2\kappa_2^2 - 1/K}.$$

(20)

Also, with probability at least $1 - e^{-2t}$, for every $f \in \mathcal{F}$

$$\text{Var}_{\hat{p}_n}(f) \leq \frac{K(1 + 2\kappa_2^2)}{K - 1} \text{Var}(f) + 32c_2^2KM^2r_n^* + \frac{4M^2t}{n}(2c_2 + K) + \frac{2K}{K - 1} \kappa_2^2.$$

(21)

See Section 4.2.5 for the proof.

Applying the upper bound on the variance (20) to (18) and noting that $\frac{2\rho}{n} \geq \frac{\kappa_1^2}{1 - 2\kappa_2^2 - 1/K}$, with probability at least $1 - e^{-3t}$

$$\mathbb{E}[f] \leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{\kappa_1^2}{1 - 2\kappa_2^2 - 1/K} \text{Var}_{\hat{p}_n}(f) + \kappa_2 \left( \sqrt{\frac{2\kappa_1^2}{1 - 2\kappa_2^2 - 1/K}} + 1 \right)}$$

$$+ \sqrt{\frac{\kappa_1^2}{1 - 2\kappa_2^2 - 1/K} \left( 32c_2^2KM^2r_n^* + \frac{4M^2t}{n}(2c_2 + K) \right)}$$

$$\leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{2\rho}{n} \text{Var}(f) + \kappa_2 \left( 2\sqrt{\frac{\rho}{n}} + 1 \right) + \frac{2\rho}{n} \left( 32c_2^2KM^2r_n^* + \frac{4M^2t}{n}(2c_2 + K) \right)}$$

13
for every $f \in F$. Finally, noting that

$$\kappa_1^2 \leq \frac{4c}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + 2\kappa_1^2 r_n^* + \frac{16c_1c_2M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right),$$

$$\frac{\rho}{2\kappa_1^2 n} \geq \frac{8c_2M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \text{ and } \sqrt{a + b} \leq \sqrt{a} + \sqrt{b} \text{ for } a, b > 0, \text{ we arrive at (19).}$$

Recall the lower bound (8) given in Theorem 1

$$\mathbb{E}[\hat{\rho}_n[f] + \sqrt{2\rho/n \text{Var}(\hat{\rho}_n(f))] \leq \sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)] + 2\frac{M\rho}{n},$$

Using this bound in the uniform empirical Bernstein’s inequality (19), with probability at least $1 - e^{-3t}$, for every $f \in F$

$$\mathbb{E}[f] \leq \sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)] + c_1 \left( 1 + 2\sqrt{\frac{2}{n}} r_n^* + 8c_1 M \sqrt{\frac{r_n^* K \rho}{n}} + \frac{\sqrt{\rho}}{n} \left( 3M \sqrt{(2c_2 + K)t} + \frac{1}{2c_1} \left( \sqrt{\rho} + \frac{2\rho}{n} \right) \right) \right).$$

Plugging in $\hat{f} = \arg\min_{f \in F} \sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)]$ and letting $K = 3, \alpha = \frac{1}{2}$ so that $c_1 = 3, c_2 = \frac{5}{2}$, we obtain (11a).

To see the bound (11b), recall that from the upper bound (8) given in Theorem 1, for every $f \in F$

$$\sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)] \leq \sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)] \leq \mathbb{E}[\hat{\rho}_{\hat{\rho}}[f] + \sqrt{\frac{2\rho}{n} \text{Var}(\hat{\rho}_n(f))} \text{ a.s.}.$$

Applying the bound (21) and Lemma A.4 to the preceding display. With probability $1 - e^{-3t}$, for every $f \in F$

$$\sup_{P: D_\phi(P|\hat{P}_n) \leq \xi} \mathbb{E}[\hat{\rho}_n(f)] \leq \mathbb{E}[\hat{\rho}_n[f] + \sqrt{\frac{K(1 + 2\kappa_1^2) 2\rho}{K - 1} \text{Var}(f)} \right.$$

$$\left. + 2\kappa_2 \sqrt{\rho K - \frac{1}{n}} + 8c_1 M \sqrt{\frac{K \rho}{n} r_n^* + \frac{2M}{n} \sqrt{2(2c_2 + K)\rho t}} \right.$$

$$\leq \mathbb{E}[f] + c_1 \sqrt{\text{Var}(f)} + \sqrt{\frac{K(1 + 2\kappa_1^2) 2\rho}{K - 1} \text{Var}(f)} \right.$$

$$\left. + \kappa_2 \left( \sqrt{\frac{K(1 + 2\kappa_1^2) 2\rho}{K - 1} + 1} + 8c_1 M \sqrt{\frac{K \rho}{n} r_n^* + \frac{2M}{n} \sqrt{2(2c_2 + K)\rho t}} \right) \right).$$

where we used $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ and the bound (21) in the first inequality and Lemma A.4 in the second inequality.

Now, note that we have $\sqrt{2\rho/n} \geq \kappa_1^2$ and $\kappa_1^2 \leq \frac{1}{4}$ for $\alpha = \frac{1}{2}, c_1 = 3$ and $c_2 = \frac{5}{2}$ by hypothesis.

Applying the upper bound on the robust objective (23) to our previous result (22), with probability at least $1 - e^{-6t}$, the right hand side of (22) is bounded by

$$\mathbb{E}[f] + 2\sqrt{5K - \frac{2\rho}{K - 1} \text{Var}(f)} + 2c_1 r_n^* \left( 1 + \sqrt{\frac{\rho}{n}} + \sqrt{\frac{K \rho}{(K - 1)n}} \right) + 16c_1 M \sqrt{\frac{K \rho}{n} r_n^*}$$

$$+ \frac{\sqrt{\rho}}{n} \left( 9M \sqrt{(2c_2 + K)t} + \frac{\sqrt{\rho}}{c_1} \left( 1 + \sqrt{\frac{\rho}{n}} \right) + \frac{16c_2\rho}{n} \sqrt{\frac{K}{K - 1} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)} \right)$$

for all $f \in F$. Plugging in the values of $c_1, c_2$ and $K = 3$, (11b) follows.
A.3 Proof of Theorem 2

Throughout this proof, we let $\text{Var}_{\hat{P}_n}(f) = \mathbb{E}_{\hat{P}_n}[f(X)^2] - \mathbb{E}_{\hat{P}_n}[f(X)]^2$ denote the empirical variance of the function $f$, and we use $\sigma^2_Q(f) = \mathbb{E}_Q[(f - \mathbb{E}_Q[f])^2]$ to denote the variance of $f$ under the distribution $Q$. Our starting point is to recall from inequality (17) in the proof of Theorem 1 that for each $f \in \mathcal{F}$, the empirical variance equality (10) holds if $n \geq \frac{4\rho M^2}{\sigma^2_Q(f)}$. As a consequence, Theorem 2 will follow if we can provide a uniform lower bound on the sample variances $s^2_n(f)$ that holds with high enough probability.

As in Section A.2 let $c_1 := (1 + \alpha), c_2 := \frac{1}{3} + \frac{1}{\alpha}$ for some $\alpha > 0$ to be chosen later. For ease of notation, define

$$
\kappa_1 = \sqrt{\frac{2c}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + c_1 \left( r_n^* + \frac{8c_2 M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right),}
$$

$$
\kappa_2 = c_1 \left( r_n^* + \frac{8c_2 M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right).
$$

From Lemma A.6 with $K = 2$, we have for all $f \in \mathcal{F}$

$$
\text{Var}_{\hat{P}_n}(f) \geq \frac{1}{2} (1 - 4\kappa_1^2) \tau - 32c_1^2 KM^2 r_n^* - \frac{4M^2 t}{n}(2c_2 + K) \quad (24)
$$

with probability at least $1 - e^{-t}$.

Note that from the hypothesis of the theorem, we have

$$
\frac{n}{48} \geq 3nr_n^* + 10M \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \quad \text{and}
$$

$$
\frac{n\tau}{16M^2} \geq 145nr_n^* + \rho + 5t + \frac{1}{M} \log \left( \log \frac{4nM^2}{t} + 1 \right).
$$

Let $\alpha = \frac{1}{2}$ so that $c_1 = 3$ and $c_2 = \frac{5}{6}$. Then,

$$
\kappa_1^2 \leq \frac{4c}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + 2c_1 r_n^* + \frac{16c_1 c_2 M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)
$$

$$
\leq \frac{60M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + 18r_n^* + 3M \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)
$$

and

$$
\kappa_2 \leq \frac{3(r_n^* + 3M \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right))}{48} \leq \frac{1}{16} \left( r_n^* + \frac{3M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right).
$$

Plugging these values into (24), we have

$$
\text{Var}_{\hat{P}_n}(f) \geq \frac{\tau}{4} - 64 \cdot 9M^2 r_n^* - \frac{16M^2 t}{n} - \frac{1}{8} \left( r_n^* + \frac{3M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right) \geq \frac{4\rho M^2}{n}
$$

which gives the final result.

A.4 Proof of Lemma 3.1

As before, let $c_1 := (1 + \alpha), c_2 := \frac{1}{3} + \frac{1}{\alpha}$ for some $\alpha > 0$ to be chosen later and let $K > 1$ be an arbitrary but fixed. From the second part of Lemma A.3 we have with probability at least $1 - e^{-t}$

$$
\{ cf : f \in \mathcal{F}, c \in [0, 1], \mathbb{E}[c^2 f^2] \leq \tau \} \subseteq \{ cf : f \in \mathcal{F}, c \in [0, 1], \mathbb{E}_{\hat{P}_n}[c^2 f^2] \leq \Delta(r) \}.
$$
where $\Delta(r) = \frac{K}{Km} \left( r + 8c_1^2Km^2r_n^* + \frac{M_2^2}{\eta}(2c_2 + K) \right)$. Letting $t = \log n$ and considering the Rademacher averages, we have

$$\psi_n(r) \leq \mathbb{E}_{\mathcal{R}_n} \left\{ cf : f \in \mathcal{F}, c \in [0, 1], \mathbb{E}_{\tilde{P}_n} |c^2f|^2 \leq \Delta(r) \right\} + \frac{M}{n}.$$

We use the standard notion of the covering number, which we now define. Let $\mathcal{V}$ be a vector space and $V \subseteq \mathcal{V}$ be any collection of vectors in $\mathcal{V}$. Let $\|\cdot\|$ be a (semi)norm on $\mathcal{V}$. We say a collection $v_1, \ldots, v_N \subseteq \mathcal{V}$ is an $\epsilon$-cover of $\mathcal{V}$ if for each $v \in \mathcal{V}$, there exists $v_i$ such that $\|v - v_i\| \leq \epsilon$. The covering number of $\mathcal{V}$ with respect to $\|\cdot\|$ is then

$$N(V, \epsilon, \|\cdot\|) := \inf \{ N \in \mathbb{N} : \text{there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\| \}.$$ 

Denote by $L_2(\tilde{P}_n)$ the $L_2$-norm with respect to the empirical measure. From a standard chaining argument (see, for example, van der Vaart and Wellner [27 Corollary 2.2.8]), we have

$$\mathbb{E}_{\mathcal{R}_n} G \leq C \sqrt{n} \int_0^{2\Delta(r_n^*)} \sqrt{\log N(G, \epsilon, L_2(\tilde{P}_n))} d\epsilon,$$

where $G := \{ cf : f \in \mathcal{F}, c \in [0, 1], \mathbb{E}_{\tilde{P}_n} |c^2f|^2 \leq \Delta(r) \}$ and $C > 0$ is a universal constant that may take different values line to line. Given an $\epsilon/2$-cover of $\mathcal{F}$ and a $\epsilon/2$-cover of $[0, 1]$, we can construct an $\epsilon$-cover of $G$. That is, we can bound

$$\log N(G, \epsilon, L_2(\tilde{P}_n)) \leq \log N \left( [0, 1], \frac{\epsilon}{2}, \|\cdot\| \right) + \log N \left( \mathbb{E}, \frac{\epsilon}{2}, L_2(\tilde{P}_n) \right).$$

Next, we use the following standard result (van der Vaart and Wellner [27 Theorem 2.6.7]) which bounds the covering number by the VC-dimension.

**Lemma A.7.** For any probability measure $Q$ on $\mathcal{X}$, we have

$$N(\mathcal{F}, M\epsilon, L_2(Q)) \leq C \mathbb{V}C(\mathcal{F})(16\epsilon)^{\mathbb{V}C(\mathcal{F}) - 1}.$$ 

Combining these bounds, we have

$$r_n^* = \psi_n(r_n^*) \leq C \sqrt{\mathbb{V}C(\mathcal{F})} \int_0^{2\Delta(r_n^*)} \sqrt{\log(M/\epsilon)} d\epsilon \leq C \sqrt{\mathbb{V}C(\mathcal{F}) \Delta(r_n^*) \log(1/\Delta(r_n^*))} \leq C \sqrt{\mathbb{V}C(\mathcal{F}) \Delta(r_n^*) \log(Kn)} \leq CM \sqrt{\mathbb{V}C(\mathcal{F})} \frac{K r_n^* \log n}{n} + CM \sqrt{\mathbb{V}C(\mathcal{F}) \log n}.$$

Solving for $r_n^*$, we obtain

$$r_n^* \leq C K M \mathbb{V}C(\mathcal{F}) \frac{\log n}{n}.$$

Letting $K \downarrow 1$, we obtain the result.

**A.5 Proof of Lemma A.7**

We proceed as in Section A.4 where instead of Lemma A.7, we use Lipschitzness of the loss function $N(\mathcal{F}, \epsilon, L_2(\tilde{P}_n)) \leq N(\Theta, \frac{\epsilon}{L}, \|\cdot\|) \leq \left( 1 + \frac{L \text{diam}(\Theta)}{\epsilon} \right)^d.$

Noting that $M = L \text{diam}(\Theta)$, we obtain the result.
We now show the claim for the empirical risk minimizer. Let \( \ell \)

\[ R(\theta^{\text{rob}}) \leq R(\theta^*) + \frac{C_3 \rho}{n}. \]

We certainly have \( \theta \)

\[ \therefore \text{we have for all } \theta \in \Theta \text{, with probability at least } \frac{1 - \delta}{\sqrt{n}} \]

of \( \delta \)

\[ \text{we have } R(\hat{\theta}^{\text{erm}}) = R(\theta^*) + n^{-\frac{1}{2}}. \]

The risk for the empirical risk minimizer, as Lemma A.8 shows, may be substantially higher; taking \( \delta = 1/\sqrt{n} \) we see that with probability at least \( 2\Phi(-\frac{n}{\sqrt{n-1}} - 2\sqrt{2}/\sqrt{\pi n}) \geq 2\Phi(-\frac{n}{\sqrt{n-1}} - n^{-\frac{1}{2}}), \)

\[ \text{we have } R(\hat{\theta}^{\text{erm}}) - R(\theta^*) \geq \delta. \]

We begin with the following

\[ \text{Claim A.1. If } \hat{S}^c \not\subset S^{2\epsilon}, \text{ then} \]

\[ \sup_{\theta \in \hat{S}^{2\epsilon}} \left\{ \Delta_n(\theta) + \sqrt{\frac{2\rho}{n}} \text{Var}_{\hat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X)) \right\} \geq \epsilon. \]  

\[ \text{Deferring the proof of the claim, let us prove the theorem. First, the growth condition (13) shows that} \]

\[ S^{2\epsilon} \subset \left\{ \theta \in \Theta : \|\theta - \pi(\theta)\|_2 \leq \left( \frac{2\epsilon}{\lambda} \right)^{\frac{1}{\gamma}} \right\} = \left\{ \theta \in \Theta : \text{dist}(\theta, S) \leq \left( \frac{2\epsilon}{\lambda} \right)^{\frac{1}{\gamma}} \right\}. \]

Therefore, we have for all \( \theta \in S^{2\epsilon} \) that

\[ \text{Var}_{\hat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X)) \leq L^2 \text{dist}(\theta, S)^2 \leq L^2 \left( \frac{2\epsilon}{\lambda} \right)^{\frac{2}{\gamma}}, \]

and so by the assumption (14) that \( \epsilon \geq \left( \frac{8L^2 \epsilon}{n}\right)^{\frac{1}{\gamma - 1}} (\frac{2}{\lambda})^{\frac{1}{\gamma - 1}} \), we have

\[ \sqrt{\frac{2\rho}{n}} \text{Var}_{\hat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X)) \leq \frac{2\rho}{n} \left( \frac{2\epsilon}{\lambda} \right)^{\frac{1}{\gamma}} \leq \frac{\epsilon}{2}. \]

In particular, if the event (26) holds then

\[ \sup_{\theta \in S^{2\epsilon}} \Delta_n(\theta) \geq \frac{\epsilon}{2}, \]

and recalling the definition (25) of \( \Delta_n \), it then follows that

\[ P \left( \hat{S}^c \not\subset S^{2\epsilon} \right) \leq P \left( \sup_{\theta \in S^{2\epsilon}} \Delta_n(\theta) \geq \frac{\epsilon}{2} \right). \]
To bound the probability (27), we use standard bounded difference and symmetrization arguments [e.g. [7] Theorem 6.5]. Letting $f(X_1, \ldots, X_n) := \sup_{\theta \in S^{2c}} \Delta_n(\theta)$, the function $f$ satisfies bounded differences:

$$\sup_{x, x' \in \mathcal{X}} |f(X_1, \ldots, X_{j-1}, x, X_{j+1}, \ldots, X_n) - f(X_1, \ldots, X_{j-1}, x', X_{j+1}, \ldots, X_n)|$$

$$\leq \sup_{x, x' \in \mathcal{X}} \sup_{\theta \in S^{2c}} \left| \frac{1}{n} (\ell(\theta; x) - \ell(\pi(\theta); x)) - \frac{1}{n} (\ell(\theta; x') - \ell(\pi(\theta); x')) \right|$$

$$\leq \frac{2L}{n} \sup_{\theta \in S^{2c}} \text{dist}(\theta, S) \leq \frac{2L}{n} \left( \frac{2c}{\lambda} \right)^{\frac{1}{2}}$$

for $j = 1, \ldots, n$. Using the standard symmetrization inequality $\mathbb{E}[\sup_{\theta \in S^{2c}} \Delta_n(\theta)] \leq 2\mathbb{E}[\mathcal{R}_n(S^{2c})]$ and the bounded differences inequality [7] Theorem 6.5], we have

$$\mathbb{P} \left( \sup_{\theta \in S^{2c}} \Delta_n(\theta) \geq 2\mathbb{E}[\mathcal{R}_n(S^{2c})] + t \right) \leq \exp \left( -\frac{nt^2}{2L^2} \left( \frac{\lambda}{2c} \right)^{\frac{1}{2}} \right)$$

for all $t \geq 0$. Letting $u = \frac{nt^2}{2L^2} \left( \frac{\lambda}{2c} \right)^{\frac{1}{2}}$ above and recalling the assumption (14) upper bounding $\mathbb{E}[\mathcal{R}_n(S^{2c})]$, we have $\mathbb{P}(\sup_{\theta \in S^{2c}} \Delta_n(\theta) \geq \frac{c}{2}) \leq e^{-u}$. The first result of the theorem follows from the bound (27).

To show the last claim of Theorem 1, note that a minor extension of standard chaining arguments (see, for example, van der Vaart and Wellner [27] Section 2.2), or the argument for Lemma 3.2, we have

$$\mathbb{E}[\mathcal{R}_n(S^{2c})] \leq CL \left( \frac{2c}{\lambda} \right)^{\frac{1}{2}} \sqrt{\frac{d}{n}}$$

for some numerical constant $C > 0$. Plugging this into the bound (14) and rearranging for $\epsilon$, we obtain the result.

**Proof of Claim A.1**

If $S^c \not\subset S^{2c}$, then certainly it is the case that there is some $\theta \in \Theta \setminus S^{2c}$ such that

$$R_n(\theta, \mathcal{P}_n) \leq \inf_{\theta \in \Theta} R_n(\theta, \mathcal{P}_n) + \epsilon \leq R_n(\pi(\theta), \mathcal{P}_n) + \epsilon.$$

Using the convexity of $R_n$, we have for all $t \in [0, 1]$ that

$$R_n(t\theta + (1-t)\pi(\theta), \mathcal{P}_n) \leq tR_n(\theta, \mathcal{P}_n) + (1-t)R_n(\pi(\theta), \mathcal{P}_n) \leq R_n(\pi(\theta), \mathcal{P}_n) + t\epsilon.$$

For all $t \in [0, 1]$, we have by definition of orthogonal projection (because the vector $\theta - \pi(\theta)$ belongs to the normal cone to $S$ and $\pi(\theta)$; cf. [13] Prop. III.5.3.3]) that $\pi(t\theta + (1-t)\pi(\theta)) = \pi(\theta)$. Thus, choosing $t$ appropriately there exists $\theta' \in \partial d S^{2c}$ with $\theta' = t\theta + (1-t)\pi(\theta)$ and $\pi(\theta') = \pi(\theta)$, and

$$R_n(\theta', \mathcal{P}_n) \leq R_n(\pi(\theta'), \mathcal{P}_n) + \epsilon.$$

Adding and subtracting the risk $R(\theta)$ and $R(\pi(\theta))$, we have that for some $\theta \in \partial d S^{2c}$ that

$$R_n(\theta, \mathcal{P}_n) - R(\theta) + R(\pi(\theta)) - R_n(\pi(\theta), \mathcal{P}_n) \leq R(\pi(\theta)) - R(\theta) + \epsilon \leq -\epsilon,$$

where we have used that $R(\theta) = R(\pi(\theta)) + 2\epsilon$ by construction. Multiplying by $-1$ on each side of the preceding display and taking suprema, we find that

$$\epsilon \leq \sup_{\theta \in S^{2c}} \{ R(\theta) - R_n(\theta, \mathcal{P}_n) - (R(\pi(\theta)) - R_n(\pi(\theta), \mathcal{P}_n)) \}$$

$$\leq \sup_{\theta \in S^{2c}} \sup_{P \in D_n(\mathcal{P}_n)} \{ R(\theta) - R(\pi) + \mathbb{E}_P [\ell(\pi(X); X) - \ell(\theta; X)] \}.$$

Applying the upper bound in inequality 2 of Theorem 1 gives the claim. \[ \Box \]

**B Proofs of Technical Lemmas**

**B.1 Proof of Lemma A.8**

Defining $N_y := \text{card} \{ i \in [n] : X_i = y \}$ for $y \in \{-1, 0, 1\}$, we immediately obtain

$$\mathbb{E}_{\tilde{P}_n} [\ell(\theta; X)] = \frac{1}{n} \left[ |N_{-1}| \theta + 1 | + N_1 | \theta - 1 | + N_0 | \theta | - (n - N_0) \right],$$
because \( N_1 + N_{-1} + N_0 = n \). In particular, we find that the empirical risk minimizer \( \theta \) satisfies
\[
\hat{\theta}_{\text{erm}} := \arg \min_{\theta \in \mathbb{R}} \mathbb{E}_{\hat{p}_n} [\ell(\theta; X)] = \begin{cases} 1 & \text{if } N_1 > N_0 + N_{-1} \\ -1 & \text{if } N_{-1} > N_0 + N_1 \\ \in [-1, 1] & \text{otherwise}. \end{cases}
\]
On the events \( N_1 > N_{-1} + N_0 \) or \( N_{-1} > N_0 + N_1 \), which are disjoint, then, we have
\[
R(\hat{\theta}_{\text{erm}}) = \delta = R(\theta^*) + \delta.
\]
Let us give a lower bound on the probability of this event. Noting that marginally \( N_1 \sim \text{Bin}(n, \frac{1-\delta}{2}) \) and using \( N_0 + N_{-1} = n - N_1 \), we have \( N_1 > N_0 + N_{-1} \) if and only if \( N_1 > \frac{n}{2} \), and we would like to lower bound
\[
\mathbb{P}\left( N_1 > \frac{n}{2} \right) = \mathbb{P}\left( \text{Bin}\left(n, \frac{1-\delta}{2}\right) > \frac{n}{2} \right) = \mathbb{P}\left( \text{Bin}\left(n, \frac{1+\delta}{2}\right) < \frac{n}{2} \right).
\]
Letting \( \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du \) denote the standard Gaussian CDF, then Zubkov and Serov [31] show that
\[
\mathbb{P}\left( N_1 \geq \frac{n}{2} \right) \geq \Phi\left( -\sqrt{2nD_{\text{kl}}\left( \frac{1}{2}, \frac{1+\delta}{2} \right)} \right)
\]
where \( D_{\text{kl}}(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \) denotes the binary KL-divergence. We have by standard bounds on the KL-divergence [24] Lemma 2.7] that \( D_{\text{kl}}(\frac{1}{2}||\frac{1+\delta}{2}) \leq \frac{\delta^2}{2(1-\delta^2)} \), so that
\[
\mathbb{P}\left( N_1 > \frac{n}{2} \text{ or } N_{-1} > \frac{n}{2} \right) \geq 2\Phi\left( -\sqrt{\frac{n\delta^2}{1-\delta^2}} \right) - 2\mathbb{P}\left( N_1 = \frac{n}{2} \right).
\]
For \( n \) odd, the final probability is 0, while for \( n \) even, we have
\[
\mathbb{P}\left( N_1 = \frac{n}{2} \right) = 2^{-n}\left( n/2 \right) (1-\delta^2)^{n/2} \leq (1-\delta^2)^{n/2} \sqrt{\frac{2}{\pi n}},
\]
where the inequality uses that \( (2n)! \leq \frac{4^n}{\sqrt{\pi n}} \) by Stirling’s approximation. Summarizing, we find that
\[
\mathbb{P}\left( N_1 > \frac{n}{2} \text{ or } N_{-1} > \frac{n}{2} \right) \geq 2\Phi\left( -\sqrt{\frac{n\delta^2}{1-\delta^2}} \right) - (1-\delta^2)^{n/2} \sqrt{\frac{8}{\pi n}}.
\]

**B.2 Proof of Lemma [A.1]**

We use two technical lemmas in the proof of this lemma.

**Lemma B.1** (Samson [23], Corollary 3). Let \( f : \mathbb{R}^n \to \mathbb{R} \) be convex and L-Lipschitz continuous with respect to the \( \ell_2 \)-norm over \([a, b]^n\), and let \( Z_1, \ldots, Z_n \) be independent random variables on \([a, b]\). Then for all \( t \geq 0 \),
\[
\mathbb{P}(f(Z_{1:n}) \geq \mathbb{E}[f(Z_{1:n})] + t) \lor \mathbb{P}(f(Z_{1:n}) \leq \mathbb{E}[f(Z_{1:n})] - t) \leq \exp\left( -\frac{t^2}{2L^2(b-a)^2} \right).
\]

**Lemma B.2.** Let \( Y_i \) be independent random variables with finite 4th moment. Then we have the following inequalities:
\[
\mathbb{E}\left[ \left( \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \right)^{\frac{\gamma}{2}} \right] \geq \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i^2] \right)^{\frac{\gamma}{2}} - \frac{1}{\sqrt{n}} \left( \frac{1}{n} \sum_{i=1}^{n} \text{Var}(Y_i^2) \right)^{\frac{1}{2}} \text{Var}(Y_i^2) \right) \right), \]
\]
and if \( |Y_i| \leq C \) for with probability 1 for all \( 1 \leq i \leq n \), then
\[
\mathbb{E}\left[ \left( \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \right)^{\frac{\gamma}{2}} \right] \geq \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i^2] \right)^{\frac{\gamma}{2}} - \min\left\{ \frac{C^2}{n}, \frac{C}{\sqrt{n}} \right\}.
\]

We defer the proof of Lemma B.2 to the end of this section in Section B.2.1. The function \( \mathbb{R}^n \ni z \mapsto \| (I - (1/n)11^T)z \|_2 \) is 1-Lipschitz with respect to the Euclidean norm, so Lemma B.1 implies
\[
P \left( \sqrt{\text{Var} \hat{P}_n(Z)} \geq \mathbb{E}[\sqrt{\text{Var} \hat{P}_n(Z)}] + t \right) \leq \mathbb{P} \left( \sqrt{\text{Var} \hat{P}_n(Z)} \leq \mathbb{E}[\sqrt{\text{Var} \hat{P}_n(Z)}] - t \right) \leq \exp \left( -\frac{nt^2}{2M^2} \right).
\]
Then
\[
\mathbb{E}[\sqrt{\text{Var} \hat{P}_n(Z)}] \leq \sqrt{\mathbb{E}[\text{Var} \hat{P}_n(Z)]} = \sqrt{1/n - 1/n\text{Var}(Z)},
\]
giving the first inequality of the lemma. For the second, let \( Y_i = Z_i - \frac{1}{n} \sum_{j=1}^n Z_j \) so that \( s_n^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 \). Applying Lemma B.2 with \( C = M \), we obtain \( \mathbb{E}[s_n]^2 \geq \sqrt{\mathbb{E}[s_n^2] - \frac{M^2}{n}} \) which gives the result.

### B.2.1 Proof of Lemma B.2

We first prove the claim (28a). To see this, we use that
\[
\inf_{\lambda \geq 0} \left\{ \frac{a^2}{2\lambda} + \frac{\lambda}{2} \right\} = \sqrt{a^2} = |a|,
\]
and taking derivatives yields that for all \( \lambda' \geq 0 \),
\[
\frac{a^2}{2\lambda} + \frac{\lambda}{2} \geq \frac{a^2}{2\lambda'} + \frac{\lambda'}{2} - \left( \frac{a^2}{2\lambda'^2} - \frac{1}{2} \right) (\lambda - \lambda').
\]
By setting \( \lambda_n = \sqrt{\frac{1}{n} \sum_{i=1}^n Y_i^2} \); we thus have for any \( \lambda \geq 0 \) that
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^\frac{1}{2} \right] = \mathbb{E} \left[ \sum_{i=1}^n Y_i^2 + \frac{\lambda_n}{2} \right] \\
\geq \mathbb{E} \left[ \sum_{i=1}^n Y_i^2 + \frac{\lambda}{2} \right] + \mathbb{E} \left[ \left( \frac{1}{2} - \sum_{i=1}^n Y_i^2 \right) (\lambda_n - \lambda) \right].
\]
Now we take \( \lambda = \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2]} \), and we apply the Cauchy-Schwarz inequality to obtain
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^\frac{1}{2} \right] \\
\geq \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right)^\frac{1}{2} - \frac{1}{2\lambda^2} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n (Y_i^2 - \mathbb{E}[Y_i^2]) \right)^2 \right]^{\frac{1}{2}} \\
= \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right)^\frac{1}{2} - \frac{1}{2\sqrt{n}\lambda^2} \left( \frac{1}{n} \sum_{i=1}^n \text{Var}(Y_i^2) \right)^\frac{1}{2} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^2 \right]^{\frac{1}{2}}
\]
where the last equality follows from independence. Using the triangle inequality, we obtain that the final expectation is bounded by \( 2\lambda = 2\sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2]} \), which gives inequality (28a). Now we give a sharper result. We have
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^\frac{1}{2} - \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right)^\frac{1}{2} \right]^2 \\
\leq 2 \sum_{i=1}^n \mathbb{E}[Y_i^2] - 2 \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right)^2 \\
= \frac{2}{\sqrt{n}} \left( \frac{1}{n} \sum_{i=1}^n \text{Var}(Y_i^2) \right)^{\frac{1}{2}}
\]
where we used the first inequality \((28a)\). Thus we also obtain the lower bound \((28b)\). The final inequality follows immediately upon noticing that \(\text{Var}(Y^2) \leq \mathbb{E}[Y^4] \leq C^2 \mathbb{E}[Y^2]\) for \(\|Y\|_\infty \leq C\).

### B.2.2 Proof of Lemma A.3

We first show the last claim for \(f \in \mathcal{F}_\text{centered} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}\). The first two result then follows by noting that \(\text{Var}(f)\) is translation invariant. Let \(L := 1 + \log \frac{n}{t}\). Define for \(l = 1, \ldots, L - 1\)

\[
\mathcal{F}_l := \left\{ f \in \mathcal{F}_\text{centered} : e^{-lr} < \mathbb{E}[f^2] \leq e^{-(l-1)r} \right\},
\]

so that \(\mathcal{F}_\text{centered} = \bigcup_{l=1}^L \mathcal{F}_l\). Let \(z > 0\) be such that \(t \leq z\). Apply the last claim of Lemma A.2 to \(\mathcal{F}_l\) for each \(l = 1, \ldots, L - 1\). Then, with probability at least \(1 - e^{-z}\), for every \(f \in \mathcal{F}_l\)

\[
\mathbb{E}[f] \leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{2ez}{n} e^{-(l-1)r}} + c_1 \mathbb{E}[\mathcal{R}_n \mathcal{F}_l] + 4M c_2 \frac{z}{n}
\]

where in the last line we have used \(e^{-lr} \leq \mathbb{E}[f^2]\) for \(f \in \mathcal{F}_l\). Similarly, apply the last claim of Lemma A.2 to \(\mathcal{F}_L\) and note that \(e^{-lr} \leq \frac{z}{n}\). With probability at least \(1 - e^{-t}\), for every \(f \in \mathcal{F}_L\)

\[
\mathbb{E}[f] \leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{2ez}{n} \mathbb{E}[f^2]} + c_1 \mathbb{E}[\mathcal{R}_n \mathcal{F}_l] + (4M c_2 + \sqrt{2}) \frac{z}{n};
\]

Taking the union bound, we have with probability at least \(1 - Le^{-z}\), for every \(f \in \mathcal{F}_L\)

\[
\mathbb{E}[f] \leq \mathbb{E}_{\hat{p}_n}[f] + \sqrt{\frac{2ez}{n} \mathbb{E}[f^2]} + c_1 \mathbb{E}[\mathcal{R}_n \mathcal{F}_l] + (4M c_2 + \sqrt{2}) \frac{z}{n}.
\]

Letting \(z = t + \log L\), we obtain the last claim.

### B.2.3 Proof of Lemma A.4

We show the last claim for \(\mathcal{F}_\text{centered}\) and note that the result follows from translation invariance of \(\text{Var}(f)\). Define the self-normalized functions in \(\mathcal{F}_\text{centered}\)

\[
\mathcal{G}_{r, \text{centered}} := \left\{ \sqrt{\frac{r}{\mathbb{E}[f^2]} f : f \in \mathcal{F}_\text{centered}} \right\}
\]

and similarly,

\[
\mathcal{G}_r := \left\{ \sqrt{\frac{r}{\mathbb{E}[f^2]} f : f \in \mathcal{F}} \right\} \subseteq \{ cf : f \in \mathcal{F}, \mathbb{E}[c^2 f^2] \leq r, c \in [0, 1] \}.
\]

From the truncation by \(r\), we have that for all \(g \in \mathcal{G}_{r, \text{centered}}\), we have \(\mathbb{E}[g^2] \leq r\). From the last part of Lemma A.3 applied to \(\mathcal{G}_{r, \text{centered}}\), we have that with probability at least \(1 - e^{-t}\)

\[
\mathbb{E}[g] \leq \mathbb{E}_{\hat{p}_n}[g] + \sqrt{\frac{2e}{n} \mathbb{E}[g^2]} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + c_1 \mathbb{E}[\mathcal{R}_n \mathcal{G}_r] + 4c_2 M \frac{t}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right). \tag{29}
\]

Note that

\[
\mathbb{E}[\mathcal{R}_n \mathcal{G}_r] \leq \mathbb{E}[\mathcal{R}_n \{ cf : f \in \mathcal{F}, \mathbb{E}[c^2 f^2] \leq r, c \in [0, 1] \}] \leq \psi_n(r) \leq \sqrt{\frac{r}{r_n}} \psi_n(r_n^*) = \sqrt{rr_n^*}
\]

from the sub-root property of \(\psi_n\) and \(\psi_n(r_n^*) = r_n^*\). Using this upper bound in \((29)\), we get

\[
\mathbb{E}[g] \leq \mathbb{E}_{\hat{p}_n}[g] + \sqrt{\frac{2e}{n} \mathbb{E}[g^2]} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + A \sqrt{r} + C \tag{30}
\]
where $A = c_1 \sqrt{\frac{r}{n}}$ and $B = \frac{4c_2 M}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right)$.

Choose $r$ to be the largest solution to $A \sqrt{r} + B = \frac{r}{2}$ for $D = \frac{1}{c_1}$. We claim that $r \geq r_n^*$ as claimed.

From elementary algebra, we indeed get

$$r_n^* \leq r \leq r_n^* + \frac{8c_2 M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right).$$

Now, for each $g \in \mathcal{G}_{r, \text{centered}}$, there exists $f \in \mathcal{F}_{\text{centered}}$ such that $g = \sqrt{\frac{r}{\mathbb{E}[f^2]}} f$. If $\mathbb{E}[f^2] \leq r$, then rescaling (30) yields

$$\mathbb{E}[f] \leq \mathbb{E}_{\hat{\rho}_n}[f] + \sqrt{\frac{2e\mathbb{E}[f^2]}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + c_1 \left( r_n^* + \frac{8c_2 M}{c_1 n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) \right) \mathbb{E}[f^2]}.$$ 

If $\mathbb{E}[f^2] > r$, rescaling $g$ yields

$$\mathbb{E}[f] \leq \mathbb{E}_{\hat{\rho}_n}[f] + \sqrt{\frac{2e\mathbb{E}[f^2]}{n} \left( t + \log \left( \log \frac{4nM^2}{t} + 1 \right) \right) + c_1 \sqrt{r \mathbb{E}[f^2]}}.$$

instead. Combining the two cases, we have the desired result.

**B.2.4 Proof of Lemma A.5**

Below, we use the below contraction principle due to Ledoux and Talagrand [15].

**Lemma B.3.** Let $\phi$ be $L$-Lipschitz, that is, $|\phi(x) - \phi(y)| \leq L \|x - y\|$. Then, for every class $\mathcal{G}$

$$\mathbb{E}_c[\mathcal{R}_n(\phi \circ \mathcal{G})] \leq L \mathbb{E}_c[\mathcal{R}_n \mathcal{G}]$$

where $\phi \circ \mathcal{G} = \{ \phi \circ f : f \in \mathcal{G} \}$.

See Ledoux and Talagrand [15, Thm 4.4] for a detailed account.

As in Section [B.2.3] define the self-normalized functions in $\mathcal{F}$

$$\mathcal{G}_r := \left\{ \sqrt{\frac{r}{\mathbb{E}[f^2]}} \sqrt{r} f : f \in \mathcal{F} \right\} \subseteq \left\{ cf : f \in \mathcal{F}, \mathbb{E}[c^2 f^2] \leq r, c \in [0, 1] \right\}.$$

Let $\mathcal{G}_r^2 = \{ g^2 : g \in \mathcal{G}_r \}$. From the truncation by $r$, we have that for all $g^2 \in \mathcal{G}_r^2$, $\text{Var}(g^2) \leq \mathbb{E}[g^2] \leq M^2 \mathbb{E}[g^2] \leq M^2 r$. From Lemma A.2 applied to $\mathcal{G}_r^2$, with probability at least $1 - e^{-t}$, for every $g \in \mathcal{G}_r$

$$\mathbb{E}[g^2] \leq \mathbb{E}_{\hat{\rho}_n}[g^2] + c_1 \mathbb{E}[\mathcal{R}_n \mathcal{G}_r^2] + M \sqrt{\frac{2rt}{n} + c_2 \frac{M^2 t}{n}} \leq \mathbb{E}_{\hat{\rho}_n}[g^2] + 2c_1 M \mathbb{E}[\mathcal{R}_n \mathcal{G}_r] + M \sqrt{\frac{2rt}{n} + c_2 \frac{M^2 t}{n}} \leq \mathbb{E}_{\hat{\rho}_n}[g^2] + 2c_1 M \mathbb{E}[\mathcal{R}_n \mathcal{G}_r] + M \sqrt{\frac{2rt}{n} + c_2 \frac{M^2 t}{n}}$$

where in step (a) we used the contraction principle Lemma B.3. In step (b), we used

$$\mathbb{E}[\mathcal{R}_n \mathcal{G}_r] \leq \mathbb{E}[\mathcal{R}_n \{ cf : f \in \mathcal{F}, \mathbb{E}[c^2 f^2] \leq r, c \in [0, 1] \}] \leq \psi_n(r) \leq \sqrt{\frac{r}{\mathbb{E}[f^2]} \psi_n(r_n^*)} = \sqrt{r_n^*}$$

from the sub-root property of $\psi_n$ and $\psi_n(r_n^*) = r_n^*$. 

22
Let \( A = 2c_1 M \sqrt{\frac{M}{n}} + M \sqrt{\frac{\kappa t}{n}} \) and \( B = \frac{\kappa M^2 t}{n} \). Choose \( r \) to be the largest solution to \( A \sqrt{r} + B = \frac{r}{K} \).

Then, we have
\[
K^2 A^2 \leq r \leq K^2 A^2 + 2KB
\]
and in particular, we have \( r \geq K^2 A^2 \geq r^*_n \). For each \( g \in \mathcal{G}_r \), there exists \( f \in \mathcal{F} \) such that
\[
g = \sqrt{\frac{1}{\mathbb{E}[f^2]/K}} f.
\]
If \( \mathbb{E}[f^2] \leq r \), rescaling the inequality (31) and using the upper bound on \( r \), we obtain
\[
\mathbb{E}[f^2] \leq \mathbb{E}_\hat{P}_n[f^2] + \frac{r}{K} \leq \mathbb{E}_\hat{P}_n[f^2] + KA^2 + 2B.
\]
If \( \mathbb{E}[f^2] \geq r \), rescaling instead yields
\[
\mathbb{E}[f^2] \leq \mathbb{E}_\hat{P}_n[f^2] + \frac{\mathbb{E}[f^2]}{K}.
\]
Combining the two cases, we obtain
\[
\mathbb{E}[f^2] \leq \frac{K}{K - 1} \mathbb{E}_\hat{P}_n[f^2] + KA^2 + 2B.
\]
Noting that \( A \leq 2 \left(4c_1^2 M^2 + 2 \frac{M^2 t}{n}\right) \) by convexity, we have the first result. The second result similarly follows by reversing the roles of \( \mathbb{E}[f] \) and \( \mathbb{E}_\hat{P}_n[f] \) in the above argument.

### B.2.5 Proof of Lemma A.6

We will show the result holds for all \( f \in \mathcal{F}_{\text{centered}} \) from which the result for \( \mathcal{F} \) will follow by translation invariance of \( \text{Var}(f) \) and \( \text{Var}_\hat{P}_n(f) \). From the last part of Lemma A.4, we have that with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F}_{\text{centered}} \)
\[
(\mathbb{E}_\hat{P}_n[f])^2 \leq \left(\kappa_1 \sqrt{\text{Var}(f)} + \kappa_2\right)^2 \leq 2 \kappa_1^2 \text{Var}(f) + 2 \kappa_2^2
\]
(32)
where we have used \( \mathbb{E}[f] = 0 \) in the first inequality and convexity in the second. Combining the bound (32) with Lemma A.5 to upper bound the variance by its empirical counterpart, we obtain with probability at least \( 1 - e^{-2t} \), for every \( f \in \mathcal{F}_{\text{centered}} \)
\[
\text{Var}(f) \leq \frac{K}{K - 1} \text{Var}_\hat{P}_n(f) + \frac{2K}{K - 1} \kappa_1^2 \text{Var}(f) + \frac{2K}{K - 1} \kappa_2^2 + 32c_1^2 KM^2 r_n^* + \frac{4M^2 t}{n} (2c_2 + K).
\]
Rearranging with respect to \( \text{Var}(f) \), we get the upper bound (20). Similarly, we have that with probability at least \( 1 - e^{-2t} \), for every \( f \in \mathcal{F}_{\text{centered}} \)
\[
\text{Var}_\hat{P}_n(f) \leq \frac{K}{K - 1} (1 + 2 \kappa_1^2) \text{Var}(f) + \frac{2K}{K - 1} \kappa_2^2 + 32c_1^2 KM^2 r_n^* + \frac{4M^2 t}{n} (2c_2 + K).
\]

### C Efficient solutions to computing the robust expectation

As a first step, we give a brief description of our (essentially standard) method for solving the robust risk problem. Our work in this paper focuses mainly on the properties of the robust objective (4) and its minimizers (6), so we only briefly describe the algorithm we use; we leave developing faster and more accurate specialized methods to further work. To solve the robust problem, we use a gradient descent-based procedure, and we focus on the case in which the empirical sampled losses \( \{\ell(\theta, X_i)\}_{i=1}^n \) have non-zero variance for all parameters \( \theta \in \Theta \), which is the case for all of our experiments.

Recall the definition of the subdifferential \( \partial f(\theta) = \{g \in \mathbb{R}^d : f(\theta') \geq f(\theta) + \langle g, \theta' - \theta \rangle \text{ for all } \theta'\} \), which is simply the gradient for differentiable functions \( f \). A standard result in convex analysis [13, Theorem VI.4.4.2] is that if the vector \( p^* \in \mathbb{R}^n_+ \) achieving the supremum in the definition (4) of the robust risk is unique, then
\[
\partial_\theta R_n(\theta; P_n) = \partial_\theta \sup_{p \in P_n} \mathbb{E}_p[\ell(\theta; X)] = \sum_{i=1}^n p_i^* \partial_\theta \ell(\theta; X_i),
\]
where the final summation is the standard Minkowski sum of sets. As this maximizing vector \( p \) is indeed unique whenever \( \text{Var} P_\theta (\ell (\theta; X)) \neq 0 \), we see that for all our problems, so long as \( \ell \) is differentiable, so too is \( R_n(\theta, P_n) \) and

\[
\nabla \theta R_n(\theta, P_n) = \sum_{i=1}^n p_i^* \nabla \ell(\theta; X_i) \quad \text{for} \quad p^* = \arg \max_{p \in P_n} \left\{ \sum_{i=1}^n p_i \ell(\theta; X_i) \right\}. \tag{33}
\]

In order to perform gradient descent on the risk \( R_n(\theta, P_n) \), then, by equation (33) we require only the computation of the worst-case distribution \( p^* \). By taking the dual of the maximization (33), this is an efficiently solvable convex problem; for completeness, we provide in the sequel a procedure for this computation that requires time \( O(n \log n + \log \frac{1}{\epsilon} \log n) \) to compute an \( \epsilon \)-accurate solution to the maximization (33). As all our examples have smooth objectives, we perform gradient descent on the robust risk \( R_n(\cdot; P_n) \), with stepsizes chosen by a backtracking (Armijo) line search [10, Chapter 9.2].

We give a detailed description of the procedure we use to compute the supremum problem (15). In particular, our procedure requires time \( O(n \log n + \log \frac{1}{\epsilon} \log n) \), where \( \epsilon \) is the desired solution accuracy. Let us reformulate this as a minimization problem in a variable \( p \in \mathbb{R}^n \) for simplicity. Then we wish to solve

\[
\text{minimize} \; p^T z \quad \text{subject to} \; \frac{1}{2n} \| np - 1 \|_2^2 \leq \rho, \; p \geq 0, \; p^T 1 = 1.
\]

We take a partial dual of this minimization problem, then maximize this dual to find the optimizing \( p \). Introducing the dual variable \( \lambda \geq 0 \) for the constraint that \( \frac{1}{2n} \| p - \frac{1}{n} 1 \|_2^2 \leq \frac{\rho}{n} \) and performing the standard min-max swap [10] (strong duality obtains for this problem because the Slater condition is satisfied by \( p = \frac{1}{n} 1 \) ) yields the maximization problem

\[
\text{maximize} \; f(\lambda) := \inf_p \left\{ \frac{\lambda}{2} \| p - \frac{1}{n} 1 \|_2^2 - \frac{\lambda p}{n} + p^T z \mid p \geq 0, \; 1^T p = 1 \right\}. \tag{34}
\]

If we can efficiently compute the infimum (34), then it is possible to binary search over \( \lambda \). Recall the standard fact [13, Chapter VI.4.4] that for a collection \( \{ f_p \}_{p \in P} \) of concave functions, if the infimum \( f(x) = \inf_{p \in P} f_p(x) \) is attained at some \( p_0 \) then any vector \( \nabla f_{p_0}(x) \) is a supergradient of \( f(x) \). Thus, letting \( p(\lambda) \) be the (unique) minimizing value of \( p \) for any \( \lambda > 0 \), the objective (34) becomes

\[
f(\lambda) = \frac{1}{2} \| p(\lambda) - \frac{1}{n} 1 \|_2^2 - \frac{\lambda p(\lambda)}{n} + (p(\lambda))^T z, \]

whose derivative with respect to \( \lambda \) (holding \( p \) fixed) is

\[
f'(\lambda) = \frac{1}{2} \| p(\lambda) - \frac{1}{n} 1 \|_2^2 - \frac{\lambda p(\lambda)}{n}.
\]

Now we use well-known results on the Euclidean projection of a vector to the probability simplex [11] to provide an efficient computation of the infimum (34). First, we assume with no loss of generality that \( z_1 \leq z_2 \leq \cdots \leq z_n \) and that \( 1^T z = 0 \), because neither of these changes the original optimization problem (as \( 1^T p = 0 \) and the objective is symmetric). Then we define the two vectors \( s, \sigma^2 \in \mathbb{R}^n \), which we use for book-keeping in the algorithm, by

\[
s_i = \sum_{j \leq i} z_j, \quad \sigma_i^2 = \sum_{j \leq i} z_j^2,
\]

and we let \( z^2 \) be the vector whose entries are \( z_j^2 \). The infimum problem (34) is equivalent to projecting the vector \( \nu(\lambda) \in \mathbb{R}^n \) defined by

\[
u_i = \frac{1}{n} - \frac{1}{\lambda} z_i
\]

onto the probability simplex. Notably [11], the projection \( p(\lambda) \) has the form \( p_i(\lambda) = (v_i - \eta)_+ \) for some \( \eta \in \mathbb{R} \), where \( \eta \) is chosen such that \( \sum_{i=1}^n p_i(\lambda) = 1 \). Finding such a value \( \eta \) is equivalent [11, Figure 1] to finding the unique index \( i \) such that

\[
\sum_{j=1}^i (v_j - v_i) < 1 \quad \text{and} \quad \sum_{j=1}^{i-1} (v_j - v_{i+1}) \geq 1,
\]

taking \( i = n \) if no such index exists (the sum \( \sum_{j=1}^n (v_j - v_i) \) is increasing in \( i \) and \( v_1 - v_n = 0 \).

Given the index \( i \), algebraic manipulations show that \( \eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{i} \sum_{j=1}^i z_j / \lambda = \frac{1}{n} - \frac{1}{i} - \frac{1}{i} s_i / \lambda \)
We summarize this discussion with pseudo-code in Figures 3 and 4, which provide a main routine and computations.

\[
\text{Figure 3: Procedure F}
\]

\[
\text{Figure 4: (Figure 4)}
\]

\[
\text{Inputs: Sorted vector } z \in \mathbb{R}^n \text{ with } 1^T z = 0, \text{ parameter } \rho > 0, \text{ solution accuracy } \epsilon
\]

\[
\text{Set } \lambda_{\min} = 0 \text{ and } \lambda_{\max} = \lambda_{\infty} = \max \{ n \| z \|_\infty, \sqrt{n/2\rho} \| z \|_2 \}
\]

\[
\text{Set } s_i = \sum_{j \leq i} z_j \text{ and } \sigma_i^2 = \sum_{j \leq i} z_j^2
\]

\[
\text{While } |\lambda_{\max} - \lambda_{\min}| > \epsilon \lambda_{\infty}
\]

\[
\text{Set } \lambda = \frac{\lambda_{\max} + \lambda_{\min}}{2}
\]

\[
\text{Set } (\eta, i) = \text{FINDSHIFT}(z, \lambda, s) \text{ } \text{// (Figure 4)}
\]

\[
\text{Set } f'(\lambda) = \frac{1}{2\lambda} \| p(\lambda) - n^{-1}1 \|_2^2 - \frac{\lambda \rho}{n} + p(\lambda)^T z
\]

\[
= \frac{1}{2} \| p(\lambda) - n^{-1}1 \|_2^2 - \frac{\rho}{n} = \frac{1}{2} \sum_{j=1}^{i} (v_j - \eta - n^{-1})^2 + \frac{1}{2} \sum_{j=i+1}^{n} \frac{1}{n^2} - \frac{\rho}{n}
\]

\[
= \frac{1}{2} \sum_{j=1}^{i} \left( \frac{1}{\lambda} z_j + \eta \right)^2 + \frac{n-i}{2n^2} - \frac{\rho}{n} = \frac{\sigma_i^2}{2\lambda^2} + \frac{i\eta^2}{2} + \frac{s_i \lambda}{\lambda} + \frac{n-i}{2n^2} - \frac{\rho}{n}
\]

\[
\text{Finding the index optimal } i \text{ can be done by a binary search, which requires } O(\log n) \text{ time, and } f'(\lambda)
\]

\[
\text{is then computable in } O(1) \text{ time using the vectors } s \text{ and } \sigma^2. \text{ It is then possible to perform a binary}
\]

\[
\text{search over } \lambda \text{ using } f'(\lambda), \text{ which requires } \log \frac{1}{\epsilon} \text{ iterations to find } \lambda \text{ within accuracy } \epsilon, \text{ from}
\]

\[
\text{which it is easy to compute } p(\lambda) \text{ via } p_i(\lambda) = (v_i - \eta)_+ = (n^{-1} - \lambda^{-1} z_i - \eta)_+.
\]

\[
\text{We summarize this discussion with pseudo-code in Figures 3 and 4, which provide a main routine and}
\]

\[
\text{sub-routine for finding the optimal vector } p. \text{ These routines show that, once provided the sorted vector}
\]

\[
\text{z with } z_1 \leq z_2 \leq \cdots \leq z_n \text{ (which requires } n \log n \text{ time to compute), we require only } O(\log \frac{1}{\epsilon} \log n)
\]

\[
\text{computations.}
\]
**Inputs:** Sorted vector $z$ with $\mathbf{1}^\top z = 0$, $\lambda > 0$, vector $s$ with $s_i = \sum_{j\leq i} z_j$

Set $i_{\text{low}} = 1$, $i_{\text{high}} = n$
If $\frac{1}{n} - \frac{2}{\lambda} \geq 0$
   Return $(\eta = 0, i = n)$
While $i_{\text{low}} \neq i_{\text{high}}$
   $i = \frac{1}{2} (i_{\text{low}} + i_{\text{high}})$
   $s_{\text{left}} = \frac{1}{\lambda} (i z_i - s_i)$  // (this is $s_{\text{left}} = \sum_{j=1}^{i} (v_j - v_i)$)
   $s_{\text{right}} = \frac{1}{\lambda} ((i + 1) z_{i+1} - s_{i+1})$  // (this is $s_{\text{right}} = \sum_{j=i+1}^{n} (v_j - v_{i+1})$)
   If $s_{\text{right}} \geq 1$ AND $s_{\text{left}} < 1$
      Set $\eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{\lambda} s_i$ and RETURN $(\eta, i)$
   Else IF $s_{\text{left}} \geq 1$
      Set $i_{\text{high}} = i - 1$
   Else
      Set $i_{\text{low}} = i + 1$
   End
Set $i = i_{\text{low}}$ and $\eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{\lambda} s_i$ and RETURN $(\eta, i)$

Figure 4: Procedure FINDSHIFT to find index $i$ and parameter $\eta$ such that, for the definition $v_i = \frac{1}{n} - \frac{1}{\lambda} z_i$, we have $v_j - \eta \geq 0$ for $j \leq i$, $v_j - \eta \leq 0$ for $j > i$, and $\sum_{j=1}^{n} (v_j - \eta)_+ = 1$. Method requires time $O(\log n)$.  

26