Journal of Machine Learning Research 19 (2018) 1-55 Submitted 12/17; Revised 11/18; Published 5/19

Variance-based Regularization with Convex Objectives

John Duchi JDUCHI@STANFORD.EDU
Department of Statistics and Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Hongseok Namkoong HNAMK@STANFORD.EDU
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, USA

Editor: Alexander Rakhlin

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.

Keywords: variance regularization, robust optimization, empirical likelihood

1. Introduction

We propose and study a new approach to risk minimization that automatically trades between bias—or approximation error—and variance—or estimation error. Let $\mathcal{X}$ be a sample space, P_0 a distribution on $\mathcal{X}$, and Θ 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

$$R(\theta) := \mathbb{E}[\ell(\theta, X)] = \int \ell(\theta, x) dP(x) \tag{1}$$

given a sample $\{X_1, \ldots, X_n\}$ drawn i.i.d. according to the distribution P. Under appropriate conditions on the loss ℓ, parameter space Θ, and random variables X, a number of researchers (Bartlett et al., 2005, 2006; Boucheron et al., 2005; Koltchinskii, 2006) have shown results of the form that with high probability,

$$R(\theta) \le \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i) + C_1 \sqrt{\frac{\mathrm{Var}(\ell(\theta, X))}{n}} + \frac{C_2}{n} \quad \text{for all } \theta \in \Theta \tag{2}$$

where C_1 and C_2 depend on the parameters of problem (1) and the desired confidence guarantee. Such bounds justify empirical risk minimization (ERM), which chooses $\widehat{\theta}_n$ 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.

Given bounds of the form above and heuristically considering the classical "bias-variance" tradeoff in estimation and statistical learning, it is natural to instead choose θ to directly minimize a quantity trading between approximation and estimation error, say of the form

$$\frac{1}{n}\sum_{i=1}^n \ell(\theta, X_i) + C\sqrt{\frac{\mathrm{Var}_{\widehat{P}_n}(\ell(\theta, X))}{n}}, \tag{3}$$

where $\mathrm{Var}_{\widehat{P}_n}$ denotes the empirical variance of its argument. Maurer and Pontil (2009) considered precisely this idea, giving a number of guarantees on the convergence and good performance of such a procedure. Unfortunately, even when the loss ℓ is convex in θ, the formulation (3) is in general 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 that provides a tractable *convex* formulation whenever the loss ℓ is convex and very closely approximates the penalized risk (3). Our approach is based on Owen's empirical likelihood (Owen, 2001) and ideas from distributionally robust optimization (Ben-Tal et al., 2009; Bertsimas et al., 2014; Ben-Tal et al., 2015). Below, we give a number of theoretical guarantees and empirical evidence for its performance.

Before summarizing our contributions, we first describe our approach. Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\phi(1) = 0$. Then the *ϕ-divergence* between distributions P and Q defined on a space $\mathcal{X}$ is

$$D_\phi(P\|Q) = \int \phi\left(\frac{dP}{dQ}\right) dQ = \int_{\mathcal{X}} \phi\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x),$$

where μ is any measure for which $P, Q \ll \mu$, and $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$. Throughout this paper, we use $\phi(t) = \frac{1}{2}(t-1)^2$, which gives the χ^2-divergence (Tsybakov, 2009). Given ϕ and a sample $X_1, \ldots, X_n$, we define the *local neighborhood of the empirical distribution with radius* ρ by

$$\mathcal{P}_n := \left\{\text{distributions } P \text{ such that } D_\phi\left(P\|\widehat{P}_n\right) \le \frac{\rho}{n}\right\},$$

where $\widehat{P}_n$ denotes the empirical distribution of the sample, and our choice of $\phi(t) = \frac{1}{2}(t-1)^2$ means that $\mathcal{P}_n$ consists of discrete distributions supported on the sample $\{X_i\}_{i=1}^n$. We then define the *robustly regularized risk*

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

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 θ whenever ℓ is convex, no matter the value of $\rho \ge 0$. Given the robust empirical risk (4), our proposed estimation procedure is to choose a parameter $\widehat{\theta}_n^{\mathrm{rob}}$ by minimizing $R_n(\theta, \mathcal{P}_n)$.

Let us now discuss a few of the properties of procedures minimizing the robust empirical risk (4). Our first main technical result, which we show in Section 2, is that for bounded loss functions, the robust risk $R_n(\theta, \mathcal{P}_n)$ is a good approximation to the variance-regularized quantity (3). That is,

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

where $\varepsilon_n(\theta) \leq 0$ and is $O_P(1/n)$ uniformly in θ. We show specifically that whenever $\ell(\theta, X)$ has suitably large variance, with high probability we have $\varepsilon_n = 0$. From variance expansions of the form (5) and empirical Bernstein inequality (2), we see that $R_n(\theta, \mathcal{P}_n)$ is a $O(1/n)$-approximation to the population risk $R(\theta)$, in contrast to the cruder $O(1/\sqrt{n})$-approximation that the empirical risk $\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)]$ provides. Based on this intuition that the robustly regularized risk $R_n(\theta; \mathcal{P}_n)$ is a tighter approximation to the population risk $R(\theta)$, we show a number of finite-sample convergence guarantees for the estimator

$$\widehat{\theta}_n^{\mathrm{rob}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \sup_P \left\{ \mathbb{E}_P[\ell(\theta, X)] : D_\phi \left(P \| \widehat{P}_n \right) \leq \frac{\rho}{n} \right\} \right\} \tag{6}$$

that are often tighter than those available for ERM (see Section 3). The above problem is a *convex* optimization problem when the original loss $\ell(\cdot; X)$ is convex and Θ is a convex set.

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

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

where C_1, C_2 are constants (which we specify) that depend on the loss ℓ and domain Θ. 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

$$R(\widehat{\theta}_n^{\mathrm{rob}}) = E[\ell(\widehat{\theta}_n^{\mathrm{rob}}; X)] \leq \inf_{\theta \in \Theta} \left\{ \mathbb{E}[\ell(\theta; X)] + 2\sqrt{\frac{2\rho}{n} \mathrm{Var}(\ell(\theta; X))} \right\} + \frac{C\rho}{n} \tag{7}$$

with high probability. When there are parameters θ with small risk $R(\theta)$ and small variance $\mathrm{Var}(\ell(\theta, X))$, this guarantees that the excess risk $R(\widehat{\theta}_n^{\mathrm{rob}}) - \inf_{\theta \in \Theta} R(\theta)$ is essentially of order $O(\rho/n)$, where ρ governs our desired confidence level. Our bounds do not require the Bernstein-type condition $\mathrm{Var}(\ell(\theta; X)) \leq MR(\theta)$ often required for ERM. Since it is often the case that M depends on global information (e.g. size of parameter space Θ), we have $\mathrm{Var}(\ell(\theta; X)) \ll MR(\theta)$, in which case the bound (7) offers a tighter guarantee than that available for the ERM solution $\widehat{\theta}_n^{\mathrm{erm}}$. In particular, we give an explicit example in Section 3.3 where our robustly regularized procedure (6) converges at rate $O(\log n/n)$ 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 (Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett et al., 2005; Boucheron et al., 2005; Bartlett et al., 2006; Boucheron et al., 2013; Koltchinskii, 2006), which is relevant when one wishes to achieve "fast rates" of convergence for statistical learning algorithms (that is, faster than the $O(1/\sqrt{n})$ guaranteed by a number of uniform convergence results (Bartlett and Mendelson, 2002; Boucheron et al., 2005, 2013)). In many cases, however, such tradeoffs require either conditions such as the Mammen and Tsybakov's noise condition (Mammen and Tsybakov, 1999; Boucheron et al., 2005) or localization results made possible by curvature conditions that relate the loss/risk and variance (Bartlett et al., 2006, 2005; Mendelson, 2014). The robust solutions (6) enjoy a different tradeoff between variance and risk than that in this literature, but essentially without conditions except compactness of Θ.

In proposing any new estimator, it is essential to understand the limits of the proposed procedure and identify situations in which its performance may be worse than existing estimators. There are indeed situations in which minimizing the robust-regularized risk (4) yields some inefficiency (for example, in classical statistical estimation problems with correctly specified model). To understand limits of the inefficiency induced by using the distributionally-robustified estimator (6), in Section 4 we study explicit finite sample properties of the robust estimator for general stochastic optimization problems, and we also provide asymptotic normality results in classical problems. There are a number of situations, based on growth conditions on the population risk R, when convergence rates faster than $1/\sqrt{n}$ (or even $1/n$) are attainable (see Shapiro et al. (2009, Chapter 5)). We show that under these conditions, the robust procedure (6) still enjoys (near-optimal) fast rates of convergence, similar to empirical risk minimization (also known as sample average approximation in the stochastic programming literature). Our study of asymptotics makes precise the asymptotic efficiency loss of the robust procedure over minimizing the standard (asymptotically optimal) empirical expectation: there is a bias term that scales as $\sqrt{\rho/n}$ in the limiting distribution of $\widehat{\theta}_n^{\rm rob}$, though its variance is optimal. o

We complement our theoretical results in Section 5, where we conclude by providing three experiments comparing empirical risk minimization strategies to robustly-regularized risk minimization (6). 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.

Related Work

The theoretical foundations of empirical risk minimization are solid (Vapnik, 1998; Bartlett and Mendelson, 2002; Boucheron et al., 2005, 2013). When the expectation of the excess loss bounds its variance, it is possible to achieve faster rates than the $O(1/\sqrt{n})$ offered by standard uniform convergence arguments (Vapnik and Chervonenkis, 1971, 1974; Bartlett et al., 2006; Koltchinskii, 2006; Boucheron et al., 2013) (see Boucheron et al. (2005, Section

5) for an overview in the case of classification, and Shapiro et al. (2009, Chapter 5.3) for more general stochastic optimization problems). Vapnik and Chervonenkis (1971, 1974) first provided such results in the context of $\{0,1\}$-valued losses for classification (see also (Anthony and Shawe-Taylor, 1993)), where the expectation of the loss always upper bounds its variance, so that if there exists a perfect classifier the convergence rates of empirical risk minimization procedures are $O(1/n)$. Mammen and Tsybakov (Mammen and Tsybakov, 1999; Tsybakov, 2004) give low noise conditions for binary classification substantially generalizing these results, which yield a spectrum of fast rates. Under related conditions, Bartlett, Jordan, and McAuliffe (2006) show similar fast rates of convergence for convex risk minimization under appropriate curvature conditions on the loss. The robust procedure (6), on the other hand, is guaranteed to provide an at most $O(1/n)$ over-estimate of the population risk and a small increase of its variance regularized population counterpart. It may be the case that the variance-regularized risk $\inf_\theta\{R(\theta) + \sqrt{\mathrm{Var}(\ell(\theta, X))/n}\}$ decreases to $R(\theta^\star)$ more slowly than $1/n$. As we note above and detail in Section 4, however, in stochastic optimization problems the variance-regularized approach (6) suffers limited degradation with respect to empirical risk minimization strategies, even under convexity and curvature properties that allow faster rates of convergence than those achievable in classical regimes, as detailed by (Shapiro et al., 2009, Chapter 5.3).

Most related to our work is that of Maurer and Pontil (2009), who propose directly regularizing empirical risk minimization by variance, providing guarantees similar to ours and giving a natural foundation off of which many of our results build. In their setting, however—as they carefully note—it is unclear how to actually solve the variance-regularized problem, as it is generally non-convex. Shivaswamy and Jebara (2010, 2011) build on this and develop an elegant approach for boosting binary classifiers based on a variance penalty applied to the exponential loss; as it is a boosting approach, their approach provides a coordinate-wise strategy for decreasing the loss, but it is not guaranteed to converge to a global minimizer and applies to classification-like problems. Our approach, handling general stochastic optimization problems, removes these obstructions.

The robust procedure (6) is based on distributionally robust optimization ideas that many researchers have developed (Ben-Tal et al., 2013; Bertsimas et al., 2014; Lam and Zhou, 2015), where the goal (as in robust optimization more broadly (Ben-Tal et al., 2009)) is to protect against all deviations from a nominal data model. In the optimization literature, there is substantial work on tractability of the problem (6), including that of Ben-Tal et al. (2013), who show that the dual of (4) often admits a standard form (such as a second-order cone problem) to which standard polynomial-time interior point methods can be applied. Namkoong and Duchi (2016) develop stochastic-gradient-like procedures for solving the problem (6), which efficiently provide low accuracy solutions (which are still sufficient for statistical tasks). Work on the statistical analysis of such procedures is nascent; Bertsimas, Gupta, and Kallus (2014) and Lam and Zhou (2015) provide confidence intervals for solution quality under various conditions, and Duchi et al. (2016) give asymptotics showing that the optimal robust risk $R_n(\widehat{\theta}_n^{\mathrm{rob}}; \mathcal{P}_n)$ is a calibrated upper confidence bound for $\inf_{\theta\in\Theta} \mathbb{E}[\ell(\theta; X)]$. They and Gotoh et al. (2015) also provide a number of asymptotic results showing relationships between the robust risk $R_n(\theta; \mathcal{P}_n)$ and variance regularization, but they do not leverage these results for guarantees on the solutions $\widehat{\theta}_n^{\mathrm{rob}}$.

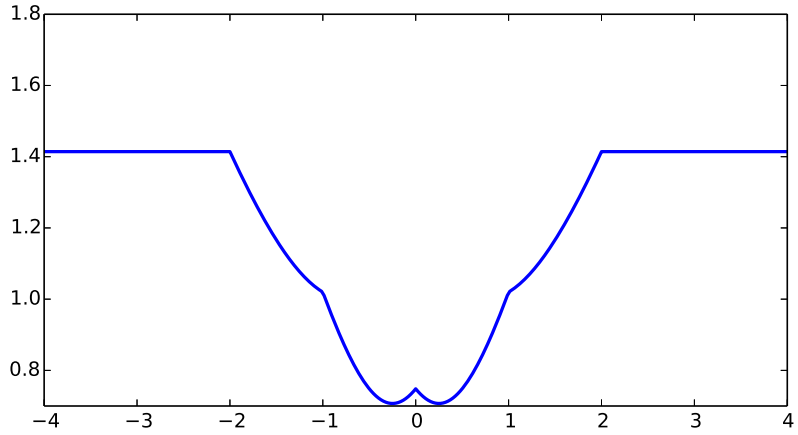


Figure 1. Plot of $\theta \mapsto \sqrt{\mathrm{Var}(\ell(\theta, X))}$ for $\ell(\theta; X) = |\theta - X|$ where $X \sim \mathsf{Uni}(\{-2, -1, 0, 1, 2\})$. The function is non-convex, with multiple local minima, inflection points, and does not grow as $\theta \to \pm\infty$.

Notation We collect our notation here. We let $\mathbb{B}$ denote a unit norm ball in $\mathbb{R}^d$, $\mathbb{B} = \{\theta \in \mathbb{R}^d : \|\theta\| \le 1\}$, where d and $\|\cdot\|$ are generally clear from context. Given sets $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^d$, we let $A + B = \{a + b : a \in A, b \in B\}$ denote Minkowski addition. For a convex function f, the subgradient set $\partial f(x)$ of f at x is $\partial f(x) = \{g : f(y) \ge f(x) + g^\top(y - x) \text{ for all } y\}$. For a function $h : \mathbb{R}^d \to \mathbb{R}$, we let h^* denote its Fenchel (convex) conjugate, $h^*(y) = \sup_x\{y^\top x - h(x)\}$. For sequences a_n, b_n, we let $a_n \lesssim b_n$ denote that there is a numerical constant $C < \infty$ such that $a_n \le C b_n$ for all n. For a sequence of random vectors $X_1, X_2, \ldots$, we let $X_n \xrightarrow{d} X_\infty$ denote that X_n converges in distribution to X_∞. For a nonegative sequence $a_1, a_2, \ldots$, we say $X_n = O_P(a_n)$ if $\lim_{c\to\infty} \sup_n \mathbb{P}(\|X_n\| \ge c a_n) = 0$, and we say $X_n = o_P(a_n)$ if $\lim_{c\to 0} \limsup_n \mathbb{P}(\|X_n\| \ge c a_n) = 0$.

2. Variance Expansion

We begin our study of the robust regularized empirical risk $R_n(\theta, \mathcal{P}_n)$ by showing that it is a good approximation to the empirical risk plus a variance term, that is, studying the variance expansion (5). Although the variance of the loss is in general non-convex (see Figure 1 for a simple example), the robust formulation (6) is a convex optimization problem for variance regularization whenever the loss function is convex (the supremum of convex functions is convex (Hiriart-Urruty and Lemaréchal, 1993, Prop. 2.1.2.)).

2.1. Variance expansion for a single variable

To gain intuition for the variance expansion that follows, we begin with a slightly simpler problem, which is to study the quadratically constrained linear maximization problem

$$\underset{p}{\text{maximize}} \ \sum_{i=1}^n p_i z_i \ \text{ subject to } \ p \in \mathcal{P}_n = \left\{p \in \mathbb{R}^n_+ : \frac{1}{2}\left\|np - \mathbf{1}\right\|_2^2 \le \rho, \langle \mathbf{1}, p\rangle = 1\right\}, \quad (8)$$

where $z \in \mathbb{R}^n$ is a vector. For simplicity, let $s_n^2 = \frac{1}{n}\|z\|_2^2 - (\overline{z})^2 = \frac{1}{n}\|z - \overline{z}\|_2^2$ denote the empirical "variance" of the vector z, where $\overline{z} = \frac{1}{n}\langle \mathbf{1}, z\rangle$ is the mean value of z. Then by introducing the variable $u = p - \frac{1}{n}\mathbf{1}$, the objective in problem (8) satisfies $\langle p, z\rangle = \overline{z} + \langle u, z\rangle = \overline{z} + \langle u, z - \overline{z}\rangle$ because $\langle u, \mathbf{1}\rangle = 0$. Thus problem (8) is equivalent to solving

$$\operatorname*{maximize}_{u \in \mathbb{R}^n} \ \overline{z} + \langle u, z - \overline{z}\rangle \quad \text{subject to} \ \ \|u\|_2^2 \le \frac{2\rho}{n^2}, \ \ \langle \mathbf{1}, u\rangle = 0, \ \ u \ge -\frac{1}{n}.$$

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

$$u_i = \frac{\sqrt{2\rho}(z_i - \overline{z})}{n\,\|z - \overline{z}\|_2} = \frac{\sqrt{2\rho}(z_i - \overline{z})}{n\sqrt{n s_n^2}}.$$

It is possible to choose such u_i while satisfying the constraint $u_i \ge -1/n$ if and only if

$$\min_{i \in [n]} \frac{\sqrt{2\rho}(z_i - \overline{z})}{\sqrt{n s_n^2}} \ge -1. \tag{9}$$

Thus, if inequality (9) holds for the vector z—that is, there is enough variance in z—we have

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

For losses $\ell(\theta, X)$ with enough variance relative to $\ell(\theta, X_i) - \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X_i)]$, that is, those satisfying inequality (9), then, we have

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

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 (9) of the exact expansion. As usual, we assume that $\phi(t) = \frac{1}{2}(t-1)^2$ in our definition of the ϕ-divergence.

Theorem 1 *Let Z be a random variable taking values in $[0, M]$. Let $\sigma^2 = \mathrm{Var}(Z)$ and $s_n^2 = \mathbb{E}_{\widehat{P}_n}[Z^2] - \mathbb{E}_{\widehat{P}_n}[Z]^2$ denote the population and sample variance of Z, respectively. Fix $\rho \ge 0$. Then*

$$\left(\sqrt{\frac{2\rho}{n}s_n^2} - \frac{2M\rho}{n}\right)_+ \le \sup_P \left\{\mathbb{E}_P[Z] : D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}\right\} - \mathbb{E}_{\widehat{P}_n}[Z] \le \sqrt{\frac{2\rho}{n}s_n^2}. \tag{10}$$

Moreover, for $n \ge \max\left\{2, \frac{M^2}{\sigma^2}\max\{8\sigma, 44\}\right\}$, with probability at least $1 - \exp\left(-\frac{3n\sigma^2}{5M^2}\right)$

$$\sup_{P : D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{2\rho}{n}s_n^2}. \tag{11}$$

See Section A for the proof of Theorem 1.

Inequality (10) and the exact expansion (11) show that, at least for bounded loss functions ℓ, 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. In the sequel, we leverage this result to provide sharp guarantees for a number of stochastic risk minimization problems.

2.2. Uniform variance expansions

We now turn to a more uniform variant of Theorem 1, which depends on familiar notions of function complexity based on Rademacher averages. For a sample $x_1, \ldots, x_n$ and i.i.d. random signs $\varepsilon_i \in \{-1, 1\}$, independent of the x_i, the empirical Rademacher complexity of the class $\mathcal{F}$ is

$$\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i)\right].$$

The *worst-case* Rademacher complexity (Srebro et al., 2010) is

$$\mathfrak{R}_n^{\sup}(\mathcal{F}) := \sup_{x_1, \ldots, x_n \in \mathcal{X}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i)\right|\right].$$

For example, when $\mathcal{F}$ is a class of functions bounded by M with VC-subgraph dimension d, we have the inequalities $\mathbb{E}[\mathfrak{R}_n(\mathcal{F})] \le \mathfrak{R}_n^{\sup}(\mathcal{F}) \lesssim M\sqrt{\frac{d}{n}}$. See van der Vaart and Wellner (1996, Chapter 2) and Bartlett and Mendelson (2002) for other bounds.

With this definition, we provide a result showing that the variance expansion (5) holds uniformly for all functions with *enough* variance.

Theorem 2 *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [0, M]$, and $M \le n$. There exists a universal constant C such that if $\tau^2 > 0$ satisfies*

$$\tau^2 \ge \frac{4\rho M^2}{n} + C\left[\mathfrak{R}_n^{\sup}(\mathcal{F})^2 \log^3 n + \frac{M^2}{n}(t + \log\log n)\right].$$

Then with probability at least $1 - 3e^{-t}$

$$\sup_{P : D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f(X)] = \mathbb{E}_{\widehat{P}_n}[f(X)] + \sqrt{\frac{2\rho}{n} \mathrm{Var}_{\widehat{P}_n}(f(X))} \tag{12}$$

for all $f \in \mathcal{F}$ such that $\mathrm{Var}(f) \ge \tau^2$.

We prove the theorem in Section B. Theorem 2 shows that the variance expansion of Theorem 1 holds uniformly for all functions f with sufficient variance. An asymptotic analogue of the equality (12) for heavier tailed random variables is also possible (Duchi et al., 2016). In the remainder of the section, we provide examples and applications of the theorem.

2.2.1. Linear and margin-based losses

Consider a standard margin-based classification problem (Bartlett and Mendelson, 2002), where we have data pairs $(x, y) \in \mathcal{X} \times \{-1, 1\}$, and $\mathcal{X} \subset \mathbb{R}^d$. Let $\Theta \subset \mathbb{R}^d$ be a norm ball of radius $r(\Theta)$, $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\| \le r\}$, and let $\|\cdot\|_*$ be the associated dual norm, assuming also that $\mathcal{X} \subset \{x \in \mathbb{R}^d \mid \|x\|_* \le r(\mathcal{X})\}$. We may then consider the standard loss minimization setting, where for some non-increasing and 1-Lipschitz loss $\ell : \mathbb{R} \to \mathbb{R}_+$, we have the risk

$$R(\theta) := \mathbb{E}\left[\ell(Y \left\langle \theta, X \right\rangle)\right],$$

so that $\ell(y \left\langle x, \theta \right\rangle)$ is the loss suffered by making prediction $\langle \theta, x \rangle$ when the label is y. By taking the function class $\mathcal{F} = \{(x, y) \mapsto \ell(y \left\langle x, \theta \right\rangle) - \ell(0) \mid \theta \in \Theta\}$, in this case, an application of the Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) implies for any $y_1, x_1, \ldots, y_n, x_n$ that

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^n \varepsilon_i \left[\ell(y_i \left\langle \theta, x_i \right\rangle) - \ell(0)\right]\right|\right] \le \mathbb{E}\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^n \varepsilon_i \left\langle \theta, x_i \right\rangle\right|\right] \le r(\Theta)\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_*\right]. \quad (13)$$

Example 1 (Euclidean norms) *In the above context, suppose that norm $\|\cdot\|$ is the standard ℓ_2 Euclidean norm so that Θ is contained in an ℓ_2-ball of radius $r(\Theta)$, and $\mathcal{X} \subset \mathbb{R}^d$ in an ℓ_2 ball of radius $r(\mathcal{X})$. Then Jensen's inequality and independence of ε_i's give the bound*

$$\mathbb{E}[\|\sum_{i=1}^n \varepsilon_i x_i\|] \le \sqrt{\mathbb{E}\sum_{j=1}^d \left(\sum_{i=1}^n \varepsilon_i x_{ij}\right)^2} \le r(\mathcal{X})\sqrt{n}.$$

Then, inequality (13) *and Theorem 1 imply that*

$$\sup_{P: D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(Y \left\langle \theta, X \right\rangle)] = \mathbb{E}_{\widehat{P}_n}[\ell(Y \left\langle \theta, X \right\rangle)] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(Y \left\langle \theta, X \right\rangle))}$$

for all θ satisfying

$$\mathrm{Var}(\ell(Y \left\langle \theta, X \right\rangle)) \ge \frac{r(\mathcal{X})^2 r(\Theta)^2}{n}\left[4\rho + C\log^3 n + Ct\right],$$

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

Example 2 (High-dimensional problems) *In high dimensional problems, the Euclidean scaling of Example 1 may be problematic, so that using ℓ_1-constraints is preferred (Bühlmann and van de Geer, 2011). Thus, taking the norm $\|\cdot\|$ in the preceding to be the ℓ_1 norm, so that $\Theta \subset \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r_1(\Theta)\}$ and $\|\cdot\|_* = \|\cdot\|_\infty$, then $\mathbb{E}[\|\sum_{i=1}^n \varepsilon_i x_i\|_\infty] \le r(\mathcal{X})\sqrt{n\log(2d)}$, where $r_\infty(\mathcal{X})$ denotes the ℓ_∞-radius of $\mathcal{X} \subset \mathbb{R}^d$. Thus, if we take the loss class $\mathcal{F} = \{\ell(\langle\theta, \cdot\rangle) - \ell(0) \mid \theta \in \Theta\}$, we obtain*

$$\mathfrak{R}_n^{\sup}(\mathcal{F}) \lesssim \sup_{x_1,\ldots,x_n\in\mathcal{X}} \frac{r_1(\Theta)}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_\infty\right] \le r_1(\Theta) r_\infty(\mathcal{X})\sqrt{\frac{\log(2d)}{n}}.$$

Then the exact variance expansion (12) *holds with probability at least $1 - e^{-t}$ uniformly over θ satisfying* $\mathrm{Var}(\ell(Y \left\langle \theta, X \right\rangle)) \ge \frac{r_1(\Theta)^2 r_\infty(\mathcal{X})^2}{n}[4\rho + C\log d \cdot \log^3 n + Ct]$.

2.2.2. Covering number guarantees

It is also possible to provide guarantees on the exact variance expansion using standard covering numbers, though careful arguments based on Rademacher complexity can be tighter. We begin by recalling the appropriate notions from approximation theory. Let $\mathcal{V}$ be a vector space and $V \subset \mathcal{V}$ be any collection of vectors in $\mathcal{V}$. Let $\|\cdot\|$ be a (semi)norm on $\mathcal{V}$. We say a collection $v_1, \ldots, v_N \subset \mathcal{V}$ is an *ϵ-cover* of $\mathcal{V}$ if for each $v \in \mathcal{V}$, there exists v_i such that $\|v - v_i\| \le \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is then

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

Now, let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, and define the $L^\infty(\mathcal{X})$ norm on f by

$$\|f - g\|_{L^\infty(\mathcal{X})} := \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

We also relax our covering number requirements to empirical ℓ_∞-covering numbers as follows. Define $\mathcal{F}(x) = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ for $x \in \mathcal{X}^n$, and define the empirical ℓ_∞-covering numbers

$$N_\infty(\mathcal{F}, \epsilon, n) = \sup_{x \in \mathcal{X}^n} N\left(\mathcal{F}(x), \epsilon, \|\cdot\|_\infty\right),$$

which bound the number of ℓ_∞-balls of radius ϵ required to cover $\mathcal{F}(x)$. Note that we always have $N_\infty(\mathcal{F}, \epsilon, n) \le N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})})$ by definition. The classical Dudley entropy integral (Dudley, 1999; van der Vaart and Wellner, 1996) shows that, if P_n denotes the point masses on $x_1, \ldots, x_n$ and $\|\cdot\|_{L^2(P_n)}$ the empirical L^2-norm on functions $f : \mathcal{X} \to [-M, M]$, then

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sup_{f \in \mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] &\lesssim \inf_{\delta \ge 0}\left\{\delta + \frac{1}{\sqrt{n}}\int_\delta^M \sqrt{\log N(\mathcal{F}, \epsilon, \|\cdot\|_{L^2(P_n)})}d\epsilon\right\} \\ &\le \inf_{\delta \ge 0}\left\{\delta + \frac{1}{\sqrt{n}}\int_\delta^M \sqrt{\log N_\infty(\mathcal{F}, \epsilon, n)}d\epsilon\right\}. \end{aligned} \tag{14}$$

Our main (essentially standard (van der Vaart and Wellner, 1996)) motivating example is that of Lipschitz loss functions for a parametric set Θ, as follows.

Example 3 *Let $\Theta \subset \mathbb{R}^d$ and assume that $\ell : \Theta \times \mathcal{X} \to [0, M]$ is L-Lipschitz in θ with respect to the ℓ_2-norm for all $x \in \mathcal{X}$, meaning that $|\ell(\theta, x) - \ell(\theta', x)| \le L\left\|\theta - \theta'\right\|_2$. Then taking $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$, any ϵ-covering $\{\theta_1, \ldots, \theta_N\}$ of Θ in ℓ_2-norm guarantees that $\min_i |\ell(\theta, x) - \ell(\theta_i, x)| \le L\epsilon$ for all θ, x. That is,*

$$N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}) \le N(\Theta, \epsilon/L, \|\cdot\|_2) \le \left(1 + \frac{\operatorname{diam}(\Theta)L}{\epsilon}\right)^d,$$

where $\operatorname{diam}(\Theta) = \sup_{\theta, \theta' \in \Theta}\left\|\theta - \theta'\right\|_2$. Thus ℓ_2-covering numbers of Θ control L^∞-covering numbers of the family $\mathcal{F}$, and we have by the entropy integral (14) that

$$\mathfrak{R}_n^{\sup}(\mathcal{F}) \lesssim \sqrt{\frac{d}{n}}\int_0^{\operatorname{diam}(\Theta)L}\sqrt{\log\frac{\operatorname{diam}(\Theta)L}{\epsilon}}d\epsilon \lesssim \operatorname{diam}(\Theta)L\sqrt{\frac{d}{n}}.$$

That is, with high probability, for all θ such that $\operatorname{Var}(\ell(\theta, X)) \ge \frac{4M^2\rho}{n} + \frac{Cd\operatorname{diam}(\Theta)^2L^2\log^3 n}{n}$, we have the exact variance expansion (12).

3. Optimization by Minimizing the Robust Loss

Based on the precise variance expansions in the preceding section, it is natural to expect that the robust solution (6) automatically trades between approximation and estimation error. This intuition is accurate, and we show that the robustly regularized objective $R_n(\theta; \mathcal{P}_n)$ overestimates the population risk $R(\theta)$ by at most $O(1/n)$. By virtue of optimizing this tighter approximation—as opposed to the usual $O(1/\sqrt{n})$-approximation given by the empirical risk $\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)]$—the robustly regularized solution (6) enjoys a number of favorable finite-sample properties, which are not always comparable to those for empirical risk minimization (ERM).

In Section 3.1, we present two versions of our main result that depend on covering numbers and discuss their consequences, and we provide an example where the robustly regularized solution $\widehat{\theta}_n^{\rm rob}$ achieves a tighter excess risk bound compared to those that a straightforward application of localized Rademacher complexities (Bartlett et al., 2005) show that the ERM solution $\widehat{\theta}_n^{\rm erm}$ achieves. As evidenced by the substantial work on Rademacher- and Gaussian-complexity and symmetrization, in some instances covering-number-based arguments do not provide the sharpest scaling (Bartlett and Mendelson, 2002; Bartlett et al., 2005; Srebro et al., 2010); thus, in Section 3.2 we present a version of our main result that depends on localized Rademacher complexities, which can allow more refined uniform concentration bounds than covering numbers. We also provide a concrete (but admittedly somewhat contrived) example where our robustly regularized procedure (6) achieves $R(\widehat{\theta}_n^{\rm rob}) - \inf_{\theta\in\Theta} R(\theta) \lesssim \frac{\log n}{n}$, while empirical risk minimization suffers $R(\widehat{\theta}_n^{\rm erm}) - \inf_{\theta\in\Theta} R(\theta) \gtrsim \frac{1}{\sqrt{n}}$, in Section 3.3. The robust "regularizer" has invariance properties other regularization procedures do not, and we mention these briefly in Section 3.4.

3.1. Covering arguments

Our first guarantee depends on the covering numbers of the function class $\mathcal{F}$ as we describe in Section 2.2.2. While we state our results abstractly, in the loss minimization setting we typically consider the function class $\mathcal{F} := \{\ell(\theta, \cdot) : \theta \in \Theta\}$ parameterized by θ. We have the following theorem, where as usual, we let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to [M_0, M_1]$ with $M = M_1 - M_0$.

Theorem 3 *Let $n \geq 8M^2/t$, $t \geq \log 12$, $\epsilon > 0$, and $\rho \geq 9t$. Then with probability at least $1 - 2(3N_\infty(\mathcal{F}, \epsilon, 2n) + 1)e^{-t}$,*

$$\mathbb{E}[f(X)] \leq \sup_{P: D_\phi(P\|\widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[f(X)] + \frac{11}{3}\frac{M\rho}{n} + \left(2 + 4\sqrt{\frac{2t}{n}}\right)\epsilon \tag{15}$$

for all $f \in \mathcal{F}$. Defining the empirical minimizer

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \left\{ \sup_P \left\{ \mathbb{E}_P[f(X)] : D_\phi(P\|\widehat{P}_n) \leq \frac{\rho}{n} \right\} \right\}$$

we have with the same probability that

$$\mathbb{E}[\widehat{f}(X)] \leq \inf_{f\in\mathcal{F}} \left\{ \mathbb{E}[f] + 2\sqrt{\frac{2\rho}{n}\mathrm{Var}(f)} \right\} + \frac{19M\rho}{3n} + \left(2 + 4\sqrt{\frac{2t}{n}}\right)\epsilon. \tag{16}$$

See Section C for a proof of the theorem. Because uniform L^∞-covering numbers upper bound empirical L^∞-covering numbers, it is immediate that covering $\mathcal{F}$ in $\|\cdot\|_{L^\infty(\mathcal{X})}$ provides an identical result.

3.1.1. Covering bounds: corollaries

We turn to a number of corollaries that expand on Theorem 3 to investigate its consequences. 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 (6) in typical empirical risk minimization scenarios.

Corollary 4 *In addition to the conditions of Theorem 3, let $\mathcal{F}$ have finite VC-dimension $\mathsf{VC}(\mathcal{F})$. Then for a numerical constant $c < \infty$, the bounds* (15) *and* (16) *hold with probability at least*

$$1 - \left(c\,\mathsf{VC}(\mathcal{F})\left(\frac{16Mne}{\epsilon}\right)^{\mathsf{VC}(\mathcal{F})-1} + 2\right)e^{-t}.$$

Proof Let $\|f\|_{L^1(Q)} := \int |f(x)| dQ(x)$ denote the L^1-norm on $\mathcal{F}$ for the probability distribution Q. Then by Theorem 2.6.7 of van der Vaart and Wellner (1996), we have

$$\sup_Q N(\mathcal{F}, \epsilon, \|\cdot\|_{L^1(Q)}) \le c\mathsf{VC}(\mathcal{F})\left(\frac{8Me}{\epsilon}\right)^{\mathsf{VC}(\mathcal{F})-1}$$

for a numerical constant c. Because $\|x\|_\infty \le \|x\|_1$, taking Q to be uniform on $x \in \mathcal{X}^{2n}$ yields $N(\mathcal{F}(x), \epsilon, \|\cdot\|_\infty) \le N(\mathcal{F}, \frac{\epsilon}{2n}, \|\cdot\|_{L^1(Q)})$. The result is immediate. ■

Next, we focus more explicitly on the estimator $\widehat{\theta}_n^{\rm rob}$ defined by minimizing the robust regularized risk (6). Let us assume that $\Theta \subset \mathbb{R}^d$, 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 $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$ in Corollary 4, we have $\mathsf{VC}(\mathcal{F}) \le d$, and we obtain the following corollary. In the corollary, recall the definition (1) of the population risk $R(\theta) = \mathbb{E}[\ell(\theta, X)]$, and the uncertainty set $\mathcal{P}_n = \{P : D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$, and that $R_n(\theta, \mathcal{P}_n) = \sup_{P\in\mathcal{P}_n} \mathbb{E}_P[\ell(\theta, X)]$. By setting $\epsilon = M/n$ in Corollary 4, we obtain the following result.

Corollary 5 *Let the conditions of the previous paragraph hold and let $\widehat{\theta}_n^{\rm rob} \in \operatorname{argmin}_{\theta\in\Theta} R_n(\theta, \mathcal{P}_n)$. Assume also that $\ell(\theta, x) \in [0, M]$ for all $\theta \in \Theta, x \in \mathcal{X}$. Then if $n \ge \rho \ge 9 \log 12$,*

$$R(\widehat{\theta}_n^{\rm rob}) \le R_n(\widehat{\theta}_n^{\rm rob}, \mathcal{P}_n) + \frac{11M\rho}{3n} + \frac{2M}{n}\left(1 + \sqrt{\frac{\rho}{n}}\right) \le \inf_{\theta\in\Theta}\left\{R(\theta) + 2\sqrt{\frac{2\rho}{n}\mathrm{Var}(\ell(\theta; X))}\right\} + \frac{11M\rho}{n}$$

with probability at least $1 - 2\exp(c_1 d \log n - c_2\rho)$, where c_i are universal constants with $c_2 \ge 1/9$.

To give an alternate concrete variant of Corollary 5 and Theorem 3, let $\Theta \subset \mathbb{R}^d$ and recall Example 3. We assume that for each $x \in \mathcal{X}$, $\inf_{\theta\in\Theta} \ell(\theta, x) = 0$ and that ℓ is L-Lipschitz in θ. If $D := \mathrm{diam}(\Theta) = \sup_{\theta,\theta'\in\Theta} \|\theta - \theta'\|_2 < \infty$, then $\ell(\theta, x) \le L\,\mathrm{diam}(\Theta)$, and

for $\delta > 0$, we define

$$\rho = \log \frac{2}{\delta} + d \log(2nDL). \tag{17}$$

Setting $t = \rho$ and $\epsilon = \frac{1}{n}$ in Theorem 3 and assuming that $\delta \lesssim 1/n$, $D \lesssim n^k$ and $L \lesssim n^k$ for a numerical constant k, choosing $\delta = \frac{1}{n}$ we obtain that with probability at least $1-\delta = 1-1/n$,

$$\mathbb{E}[\ell(\widehat{\theta}_n^{\rm rob}; X)] = R(\widehat{\theta}_n^{\rm rob}) \leq \inf_{\theta \in \Theta} \left\{ R(\theta) + C \sqrt{\frac{d \operatorname{Var}(\ell(\theta, X))}{n} \log n} \right\} + C \frac{dLD \log n}{n} \tag{18}$$

where C is a numerical constant.

3.1.2. Examples and heuristic discussion

Unpacking Theorem 3, the first result (15) (and its Corollary 5) provides a high-probability guarantee that the true expectation $\mathbb{E}[\widehat{f}]$ cannot be more than $O(1/n)$ worse than its robustly-regularized empirical counterpart. The second result (16) (and inequality (18)) guarantees convergence of the empirical minimizer to a parameter with risk at most $O(\log n/n)$ larger than the best possible variance-corrected risk.

To illustriate how variance regularization can yield tighter guarantees than empirical risk minimization by optimizing a $O(1/n)$ upper bound on the risk, we now compare the second bound (16) with an analogous result for empirical risk minimization (ERM). We first give a heuristic version, making it more precise in a coming example. For the ERM solution $\widehat{\theta}_n^{\rm erm} \in \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)]$, one common assumption is an upper bound of the variance by the risk; for example, when the losses take values in $[0, M]$, one has $\operatorname{Var}(\ell(\theta, X)) \leq MR(\theta)$. In such cases, there is typically some complexity measure $\mathfrak{Comp}_n$ associated with the class of functions being learned, and it is possible to achieve bounds of the form

$$R(\widehat{\theta}_n^{\rm erm}) \leq R(\theta^\star) + C \sqrt{\frac{\mathfrak{Comp}_n MR(\theta^\star)}{n}} + C \frac{\mathfrak{Comp}_n M}{n} \tag{19}$$

where $\theta^\star \in \operatorname{argmin}_{\theta \in \Theta} R(\theta)$, a type of result common for bounded nonnegative losses (Boucheron et al., 2005; Vapnik and Chervonenkis, 1971; Vapnik, 1998). For example, for classes of functions of VC-dimension d, we typically have $\mathfrak{Comp}_n \lesssim d \log \frac{n}{d}$. In this caricature, when $\operatorname{Var}(\ell(\theta^\star, X)) \ll MR(\theta^\star)$ and $\rho \gtrsim \mathfrak{Comp}_n$, the optimality guarantee (16) for variance regularization can be tighter than its ERM counterpart (19). This bound is certainly not always sharp, but yields minimax optimal rates in some cases.

Example 4 (Well-specified least-absolute-deviation regression) *For the least-absolute-deviation (LAD) regression, we compare rates of convergence for the ERM solution given by the localized Rademacher complexity against those for the robust solution. Let $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$, where $X \in \{x \in \mathbb{R}^d \mid \|x\|_2 \leq L\}$, and let $D := \operatorname{diam}(\Theta)$ be the ℓ_2-diameter of Θ. The LAD loss is $\ell(\theta; (x, y)) := |y - \langle \theta, x \rangle|$, where we assume that $Y = \langle \theta^\star, X \rangle + \epsilon$ for some $\theta^\star \in \Theta$, and random noise $\epsilon \in [-B, B]$ independent of X. We then have the global bound $\ell(\theta; (X, Y)) \leq DL + B =: M$. Suppose for simplicity that ϵ is uniform on $[-B, B]$; then $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ and $R(\theta^\star) = \mathbb{E}[\ell(\theta^\star; Z)] = \frac{1}{2}B$. In this case,*

$$\operatorname{Var}\left(\ell(\theta^\star; Z)\right) = \frac{B^2}{12} \leq \frac{1}{2}(DL + B)B = M\mathbb{E}[\ell(\theta^\star; Z)] = MR(\theta^\star).$$

Using that the loss is 1*-Lipschitz, the* L^∞ *covering numbers for the set of functions* $\mathcal{F} := \{f_\theta(x,y) = |\langle\theta, x\rangle - y| \mid \theta \in \Theta\}$ *satisfy* $\log N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}) \lesssim d \log \frac{DL}{\epsilon}$*, and so applying the bound* (18) *for the robustly regularized solution* $\widehat{\theta}_n^{\rm rob}$ *with* $\epsilon = DL/n$*, we obtain*

$$R(\widehat{\theta}_n^{\rm rob}) \le R(\theta^\star) + C\sqrt{\frac{d\log n}{n}B^2} + C\frac{d(LD+B)\log n}{n}$$

with probability at least $1 - 1/n$*. On the other hand, even an "optimistic" (but naive) ERM bound, achieved by taking* $\mathfrak{Comp}_n \lesssim 1$ *in the bound* (19)*, yields*

$$R(\widehat{\theta}_n^{\rm erm}) \le R(\theta^\star) + C\sqrt{\frac{\log n}{n}(BDL + B^2)} + C\frac{(LD+B)\log n}{n}$$

with probability at least $1-1/n$*. We see that leading term for the robustly regularized solution* $\widehat{\theta}_n^{\rm rob}$ *only depends on the noise-level* B^2 *while the corresponding term for the ERM solution* $\widehat{\theta}_n^{\rm erm}$ *depends on global information like the size of the parameter space* D*, and a uniform bound over covariates* L*. For typical VC and other* d*-dimensional classes, the bound* $\mathfrak{Comp}_n$ *scales linearly in* d *(cf. (Bartlett et al., 2005, Corollary 3.7), in which case the bound* (19) *scales as* $R(\theta^\star) + C\sqrt{d(BDL+B^2)\log n/n} + O(\log n/n)$*, which is worse.*

Example 5 (A hard median estimation problem) *To give a bit more insight into the behavior of the robust estimator, consider the simple 1-dimensional median problem, where* $\ell(\theta; x) = |\theta - x|$*, and assume that* $x \in \{-B, B\}$ *with* $\mathbb{P}(X = B) = \frac{1+\delta}{2}$ *for some* $\delta > 0$*, so that* $\theta^\star = \operatorname{argmin} R(\theta) = B$ *and* $R(\theta^\star) = (1-\delta)B$*. In this case, taking* $\theta_0 = 0$ *yields* $\mathrm{Var}(\ell(\theta; X)) = 0$ *and* $R(\theta_0) - R(\theta^\star) = \delta B$*. For* δ *small (on the order of* $1/\sqrt{n}$*), with constant probability the empirical risk minimizer is* $\widehat{\theta}_n^{\rm erm} = -B$*, yielding risk* $R(\widehat{\theta}_n^{\rm erm}) - R(\theta^\star) = 2\delta B$*. On the other hand, with high probability* $\widehat{\theta}_n^{\rm rob} \ge 0$ *(because* $\mathrm{Var}(\ell(\theta_0; X)) = 0$ *as* $\ell(0; X) \equiv B$*), and so* $R(\widehat{\theta}_n^{\rm rob}) - R(\theta^\star) \le \delta B$*. This gap is of course small, but it shows that the robust solution is more conservative: it chooses* $\widehat{\theta}_n^{\rm rob}$ *so that large losses (of scale* $2B$*) are less frequent.*

When the population problem is "easy", it is often possible to achieve faster rates of convergence than the usual $O(1/\sqrt{n})$ rate. The simplest scenario where this occurs is if the problem is realizable $R(\theta^\star) = 0$, in which case $\widehat{\theta}_n^{\rm erm}$ has excess risk of the order $O(\log n/n)$; see the bound (19). The robustly regularized solution $\widehat{\theta}_n^{\rm rob}$ enjoys the same faster rates of convergence under the more general condition that $\mathrm{Var}(\ell(\theta^\star; X))$ is small. As a concrete instance of this, let $\ell(\theta; X) \in [0, M]$ and assume that $\ell(\theta; X)$ satisfies the conditions of the first part of Example 3, and let the problem be realizable $R(\theta^\star) = 0$. Since $\mathrm{Var}(\ell(\theta; X)) \le MR(\theta)$, we have from the bounds (18) and (19) that

$$R(\widehat{\theta}_n^{\rm erm}) \le \frac{CdDL\log n}{n} \quad \text{and} \quad R(\widehat{\theta}_n^{\rm rob}) \le \frac{CdDL\log n}{n}.$$

For example, $\mathrm{Var}(\ell(\theta; X)) = 0$ allows for the existence of some $\theta_0 \in \Theta$ such that $\ell(\theta_0; X) < \ell(\theta^\star; X)$ with positive probability.

3.2. Localized Rademacher Complexity

A somewhat more sophisticated approach to concentration inequalities and generalization bounds is based on localization ideas, motivated by the fact that near the optimum of an empirical risk, the complexity of the function class may be smaller than over the entire (global) class (van der Vaart and Wellner, 1996; Bartlett et al., 2005). With this in mind, we now present a refined version of Theorem 3 that depends on localized Rademacher averages.

The starting point for this approach is a notion of localized Rademacher complexity (we give a slightly less general notion than Bartlett et al. (2005), as it is sufficient for our derivations). For a function class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, the localized Rademacher complexity at level r is

$$\mathbb{E}\left[\mathfrak{R}_n\left(\left\{cf \mid f \in \mathcal{F}, c \in [0,1], \mathbb{E}[c^2 f^2 \le r]\right\}\right)\right].$$

In addition, we require a few analytic notions, beginning with *sub-root* functions, where we recall (Bartlett et al., 2005) that 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 $r > 0$. Any non-constant sub-root function ψ is continuous and has a unique positive fixed point $r^\star = \psi(r^\star)$, where $r \ge \psi(r)$ for all $r \ge r^\star$. Lastly, we consider upper bounds $\psi_n : \mathbb{R}_+ \to \mathbb{R}_+$ on the localized Rademacher complexity satisfying

$$\psi_n(r) \ge \mathbb{E}[\mathfrak{R}_n(\{cf : f \in \mathcal{F}, c \in [0,1], \mathbb{E}[c^2 f^2] \le r\})], \tag{20}$$

where ψ_n is sub-root. (The localized Rademacher complexity itself is sub-root.) Roots of ψ_n play a fundamental role in providing uniform convergence guarantees, and Bartlett et al. (2005) and Koltchinskii (2006) provide careful analyses of localized Rademacher complexities, with typical results as follows. For a class of functions f with range bounded by 1, for any root $r_n^\star$ of ψ_n, with probability at least $1 - e^{-t}$ we have

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \frac{1}{\eta}\mathbb{E}_{\widehat{P}_n}[f] + C(1+\eta)\left(r_n^\star + \frac{1}{n}\right) + \frac{t}{n} \quad \text{for all } f \in \mathcal{F} \text{ and } \eta \ge 0.$$

As an example, when $\mathcal{F}$ is a bounded VC-class, we have $r_n^\star \asymp \frac{\mathsf{VC}(\mathcal{F})\log(n/\mathsf{VC}(\mathcal{F}))}{n}$ (Bartlett et al., 2005, Corollary 3.7).

With this motivation, we have the following theorem.

Theorem 6 *For $M \ge 1$, let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to [0, M]$, let ψ_n be a sub-root function bounding the localized complexity* (20)*, and let $r_n^\star \ge \psi_n(r_n^\star)$. Let $t > 0$ be arbitrary and assume that ρ satisfies*

$$\frac{\rho}{n} \ge 8\left(\frac{45M}{n}\left(t + \log\left\lceil \log\frac{n}{t}\right\rceil\right) + 18 r_n^\star\right). \tag{21}$$

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

$$\mathbb{E}[f] \le \left(1 + 2\sqrt{\frac{2\rho}{n}}\right) \sup_{P : D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f] + \left(13 + 4\sqrt{\frac{2\rho}{n}}\right)\frac{M\rho}{n} \quad \textit{for all } f \in \mathcal{F}. \tag{22}$$

Additionally, if $\widehat{f}$ *minimizes* $\sup_{P:D_\phi(P\|\widehat{P}_n)\le\rho/n}\mathbb{E}_P[f]$, *then with probability at least* $1-3e^{-t}$,

$$\mathbb{E}[\widehat{f}] \le \left(1+2\sqrt{\frac{2\rho}{n}}\right)\inf_{f\in\mathcal{F}}\left(\mathbb{E}[f]+\sqrt{\frac{91\rho}{45n}\mathrm{Var}(f)}\right)+\left(14+6\sqrt{\frac{2\rho}{n}}\right)\frac{M(3\rho+t)}{n}. \tag{23}$$

We provide the proof of Theorem 6 in Appendix D. It builds off of and parallels many of the techniques developed by Bartlett, Bousquet, and Mendelson (2005), but we require a bit of care to develop the precise variance bounds we provide.

Let us consider the additional $\sqrt{\frac{\rho}{n}}$ factors in Theorem 6 (as compared to Theorem 3). In general, these terms are negligible to the extent that the variance of f dominates the first moment of the function f—heuristically, in situations in which we expect penalizing the variance to improve performance. Let us make this more precise in a regime where n is large. Letting $f\in\mathcal{F}$, we see that we have the inequality

$$(1+\sqrt{\rho/n})\left(\mathbb{E}[f]+\sqrt{\frac{\rho}{n}\mathrm{Var}(f)}\right)\le\mathbb{E}[f]+C\sqrt{\frac{\rho}{n}\mathrm{Var}(f)}$$

(for a constant $C>1+\sqrt{\rho/n}$) if and only if $(C-1-\sqrt{\rho/n})^2\mathrm{Var}(f)\ge\mathbb{E}[f]^2$. Equivalently, as n gets large, this occurs roughly when $\mathbb{E}[f^2]\ge\frac{C^2-2C+2}{C^2-2C+1}\mathbb{E}[f]^2$, which holds for large enough C whenever $\mathrm{Var}(f)>0$.

In some scenarios, we can obtain substantially tighter bounds by using localized Rademacher averages instead of the covering number arguments considered in Section 3.1. (Recall also the discussion following Theorem 2.) To illustriate this point, we consider the case where $\mathcal{F}$ is a bounded subset of a reproducing kernel Hilbert space generated by some sufficiently nice kernel K; even for the Gaussian kernel $K(x,z)=\exp(-\frac{1}{2}\|x-z\|^2)$, log covering numbers for such function spaces grow at least exponentially in the dimension (Zhou, 2003; Kühn, 2011).

Example 6 (Reproducing kernels and least-absolute-deviation regression) *We now give an example using a non-parametric class of functionals in which covering number arguments do not apply, as the covering numbers of the associated classes are too large. Let* $\mathcal{H}$ *be a reproducing kernel Hilbert space (RKHS) with norm* $\|\cdot\|_{\mathcal{H}}$ *and associated kernel (representer of evaluation)* $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$. *Letting* P *be a distribution on* $\mathcal{X}$, *Mercer's theorem (e.g. Cristianini and Shawe-Taylor, 2004) implies that the integral operator* $T_K:L^2(\mathcal{X},P)\to L^2(\mathcal{X},P)$ *defined by* $T_K(f)(x)=\int K(x,z)dP(z)$ *is compact, and* $K(x,x')=\sum_{j=1}^\infty\lambda_j\phi_j(x)\phi_j(z)$ *where* λ_j *are the eigenvalues of* T *in decreasing order and* ϕ_j *form an orthonormal decomposition of* $L^2(\mathcal{X},P)$.

Consider now the least absolute deviation (LAD) loss function $\ell(h;x,y)=|h(x)-y|$, *defined for* $h\in\mathcal{H}$, *and let* $\mathbb{B}_{\mathcal{H}}$ *be the unit* $\|\cdot\|_{\mathcal{H}}$*-ball of* $\mathcal{H}$. *Assume additionally that the model is well-specified, and that* $y=h^\star(x)+\xi$ *for some random variable* ξ *with* $\mathbb{E}[\xi\mid X]=0$, $\mathbb{E}[\xi^2]\le\sigma^2$, *and* $h^\star\in\mathbb{B}_{\mathcal{H}}$. *Let the function class*

$$\{\ell\circ\mathcal{H}\}_{\le r}:=\left\{(x,y)\mapsto c\ell(h(x),y)\mid c\in[0,1],c^2\mathbb{E}[\ell(h(X),Y)^2]\le r\right\}.$$

Based on inequality (20), *we consider the localized complexity*

$$\mathfrak{R}_n(\{\ell\circ\mathcal{H}\}_{\le r})=\mathbb{E}\left[\frac{1}{n}\sup_{h\in\mathbb{B}_{\mathcal{H}},c\in[0,1]}\sum\varepsilon_i c\ell(h(x_i),y_i)\mid\mathbb{E}[\ell(h(X),Y)^2]\le r/c^2\right].$$

We claim that

$$\mathfrak{R}_n(\{\ell \circ \mathcal{H}\}_{\le r}) \lesssim \sqrt{r/n} + \left(\frac{1}{n}\sum_{j=1}^{\infty} \min\{\lambda_j, r\}\right)^{\frac{1}{2}}. \tag{24}$$

As this claim is not central to our development—but does show a slightly different localization result based on Gaussian comparison inequalities than available, for example, in Mendelson (2003)—we provide its proof in Appendix G.1.

Let us use inequality (24)*. To apply Theorem 3, we must find a bound on the fixed point of the localized complexity. To give this bound, we require some knowledge on the eigenvalues λ_j, for which there exists a body of work. For example (Mendelson, 2003), the Gaussian kernel $K(x, x') = \exp(-\frac{1}{2}\|x - x'\|_2^2)$ generates a class of smooth functions for which the eigenvalues λ_j decay exponentially, as $\lambda_j \lesssim e^{-j^2}$. Kernel operators underlying Sobolev spaces with different smoothness orders (Birman and Solomjak, 1967; Gu, 2002) typically have eigenvalues scaling as $\lambda_j \lesssim j^{-2\alpha}$ for some $\alpha > \frac{1}{2}$. As a concrete example, the first-order Sobolev (min) kernel $K(x, x') = 1 + \min\{x, x'\}$ generates an RKHS of Lipschitz functions with $\alpha = 1$. In the former case of $\lambda_j \lesssim e^{-j^2}$, $r_n^\star = \frac{\sqrt{\log n}}{n}$*

$$\left(\frac{1}{n}\sum_{j=1}^{\infty} \min\left\{e^{-j^2}, \frac{\log n}{n}\right\}\right)^{\frac{1}{2}} \approx \left(\frac{1}{n}\sum_{j=1}^{\sqrt{\log n}} \frac{\sqrt{\log n}}{n} + \frac{1}{n}\int_{\sqrt{\log n}}^{\infty} e^{-t^2} dt\right)^{\frac{1}{2}} \lesssim \frac{\sqrt{\log n}}{n} = r_n^\star.$$

In the latter case of polynomially decaying eigenvalues $\lambda_j \lesssim j^{-2\alpha}$, we have $j^{-2\alpha} = r$ when $r^{-\frac{1}{2\alpha}} = j$, so

$$\sum_{j=1}^{\infty} \min\{j^{-2\alpha}, r\} \approx r^{\frac{2\alpha-1}{2\alpha}} + \int_{r^{-1/2\alpha}}^{\infty} t^{-2\alpha} dt \asymp r^{\frac{2\alpha-1}{2\alpha}}.$$

Solving for $nr = r^{\frac{2\alpha-1}{2\alpha}}$, we find the fixed point $(r_n^\star)^{\frac{2\alpha-1}{4\alpha}} = r_n^\star\sqrt{n}$ yields $r_n^\star = n^{-\frac{2\alpha}{2\alpha+1}}$.

Ignoring constants, the above analysis shows that in the case that the kernel eigenvalues scale as $\lambda_j \lesssim e^{-j^2}$, as soon as $\rho \gtrsim \sqrt{\log n}$ we have

$$\mathbb{E}[\ell(h(X), Y)] \le (1+2\sqrt{2\rho/n})\left(\mathbb{E}_{\widehat{P}_n}[\ell(h(X), Y)] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(h(X), Y))}\right) + \frac{C\rho}{n} \text{ for all } h \in \mathbb{B}_{\mathcal{H}}$$

with high probability. In the case of polynomial eigenvalues, if $\widehat{h}$ minimizes the robust empirical loss $\sup_{P: D_\phi(P\|\widehat{P}_n) \le \rho/n} \mathbb{E}_P[\ell(h(X), Y)]$ and $\rho \asymp n^{1-\frac{2\alpha}{2\alpha+1}}$, then

$$\mathbb{E}\left[\ell(\widehat{h}(X), Y)\right] \le \left(1 + Cn^{-\frac{\alpha}{2\alpha+1}}\right) \inf_{h \in \mathbb{B}_{\mathcal{H}}} \left(\mathbb{E}[\ell(h(X), Y)] + Cn^{-\frac{\alpha}{2\alpha+1}}\sqrt{\mathrm{Var}(\ell(h(X), Y))}\right) + Cn^{-\frac{2\alpha}{2\alpha+1}}.$$

This rate of convergence holds without any assumptions on the smoothness of the distribution of the noise ξ.

3.3. Beating empirical risk minimization

We now provide a concrete example where the robustly regularized estimator $\widehat{\theta}_n^{\mathrm{rob}}$ exhibits a substantial performance gap over empirical risk minimization. In the sequel, we bound the

performance degradation to show that the formulation (6) in general loses little over empirical risk minimization. For intuition in this section, consider the (admittedly contrived) setting in which we replace the loss $\ell(\theta, X)$ with $\ell(\theta, X) - \ell(\theta^\star, X)$, where $\theta^\star \in \operatorname{argmin}_{\theta\in\Theta} R(\theta)$. Then in this case, by taking $\theta = \theta^\star$ in Corollary 5, we have $R(\widehat{\theta}_n^{\rm rob}) \le R(\theta^\star) + O(1/n)$ with high probability. More broadly, we expect the robustly regularized 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 concrete 1-dimensional example—estimating the median of a discrete distribution supported on $\mathcal{X} = \{-1, 0, 1\}$—in which the robustly 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

$$P(X = 1) = \frac{1-\delta}{2}, \quad P(X = -1) = \frac{1-\delta}{2}, \quad P(X = 0) = \delta. \tag{25}$$

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 $\theta^\star := \operatorname{argmin}_\theta R(\theta) = 0$, which satisfies $R(\theta^\star) = 0$. (Note also that $\ell(\theta, x) = \ell(\theta, x) - \ell(\theta^\star, x)$.) Without loss of generality, we assume that $\Theta = [-1, 1]$ in this problem.

Now, consider a sample $X_1, \ldots, X_n$ drawn i.i.d. from the distribution P, let $\widehat{P}_n$ denote its empirical distribution, and define the empirical risk minimizer

$$\widehat{\theta}_n^{\rm erm} := \operatorname*{argmin}_{\theta\in\mathbb{R}} \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)] = \operatorname*{argmin}_{\theta\in[-1,1]} \mathbb{E}_{\widehat{P}_n}[|\theta - X|].$$

If too many of the observations satisfy $X_i = 1$ or too many satisfy $X_i = -1$, then $\widehat{\theta}_n^{\rm erm}$ will be either 1 or -1; for small δ, such events become reasonably probable, as the following lemma makes precise. In the lemma, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt$ denotes the standard Gaussian CDF. (See Section G.2 for a proof.)

Lemma 7 *Let the loss $\ell(\theta; x) = |\theta - x| - |x|$, $\delta \in [0, 1]$, and X follow the distribution* (25)*. Then $R(\widehat{\theta}_n^{\rm erm}) - R(\theta^\star) \ge \delta$ with probability at least*

$$2\Phi\left(-\sqrt{\frac{n\delta^2}{1-\delta^2}}\right) - (1-\delta^2)^{\frac{n}{2}}\sqrt{\frac{8}{\pi n}}.$$

On the other hand, we certainly have $\ell(\theta^\star; x) = 0$ for all $x \in \mathcal{X}$, so that $\operatorname{Var}(\ell(\theta^\star; X)) = 0$. Now, consider the bound in Theorem 3. We see that $\log N(\{\ell(\theta, \cdot) : \theta \in \Theta\}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}) \le 2\log\frac{1}{\epsilon}$, and taking $\epsilon = \frac{1}{n}$, we have that if $\widehat{\theta}_n^{\rm rob} \in \operatorname{argmin}_{\theta\in\Theta} R_n(\theta, \mathcal{P}_n)$, then

$$R(\widehat{\theta}_n^{\rm rob}) \le R(\theta^\star) + \frac{15\rho}{n} \quad \text{with probability } \ge 1 - 4\exp\left(2\log n - \rho\right).$$

In particular, taking $\rho = 3 \log n$, we see that

$$R(\widehat{\theta}_n^{\rm rob}) \le R(\theta^\star) + \frac{45 \log n}{n} \quad \text{with probability at least } 1 - \frac{4}{n}.$$

The risk for the empirical risk minimizer, as Lemma 7 shows, may be substantially higher; taking $\delta = 1/\sqrt{n}$ we see that with probability at least $2\Phi(-\sqrt{\frac{n}{n-1}}) - 2\sqrt{2}/\sqrt{\pi e n} \ge 2\Phi(-\sqrt{\frac{n}{n-1}}) - n^{-\frac{1}{2}}$,

$$R(\widehat{\theta}_n^{\rm erm}) \ge R(\theta^\star) + n^{-\frac{1}{2}}.$$

(For $n \ge 20$, the probability of this event is $\ge .088$.) For this (specially constructed) example, there is a gap of nearly $n^{\frac{1}{2}}$ in order of convergence.

3.4. Invariance properties

The robust regularization (4) technique enjoys a number of invariance properties. Standard regularization techniques (such as ℓ_1- and ℓ_2-regularization), which generally regularize a parameter toward a particular point in the parameter space, do not. While we leave deeper discussion of these issues to future work, we make two observations, which apply when $\Theta = \mathbb{R}^d$ is unconstrained. Throughout, we let $\widehat{\theta}_n^{\rm rob} \in \argmin_\theta R_n(\theta, \mathcal{P}_n)$ denote the robustly regularized empirical solution.

First, consider a location estimation problem in which we wish to estimate the minimizer of the expectation of a loss of the form $\ell(\theta, X) = h(\theta - X)$, where $h : \mathbb{R}^d \to \mathbb{R}$ is convex and symmetric about zero. Then the robust solution is by inspection shift invariant, as $\ell(\theta + c, X + c) = \ell(\theta, X)$ for any vector $c \in \mathbb{R}^d$. Concretely, in the example of the previous section, ℓ_1- or ℓ_2-regularization achieve better convergence guarantees than ERM does, but if we shift all data $x \mapsto x + c$, then non-invariant regularization techniques lose efficiency (while the robust regularization technique does not). Second, we may consider a generalized linear modeling problem, in which data comes in pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\ell(\theta, (x, y)) = h(y, \theta^\top x)$ for a function $h : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ that is convex in its second argument. Then $\widehat{\theta}_n^{\rm rob}$ is invariant to invertible linear transformations, in the sense that for any invertible $A \in \mathbb{R}^{d \times d}$,

$$\argmin_\theta \left\{ \sup_{P : D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta, (X, Y))] \right\} = \argmin_\theta \left\{ \sup_{P : D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(A^{-1}\theta, (AX, Y))] \right\} = \widehat{\theta}_n^{\rm rob}.$$

Our results in this section do not precisely apply as we require unbounded θ, however, the next section shows that localization approaches can address this.

4. Robust regularization cannot be too bad

The previous two sections provide guarantees on the performance of the robust regularized estimator (6), it does not—cannot—dominate classical approaches based on empirical risk minimization (also known as sample average approximation in the stochastic optimization literature), though it can improve on them in some cases. For example, with a correctly specified linear regression model with gaussian noise, least-squares—empirical risk minimization with the loss $\ell(\theta, (x, y)) = \frac{1}{2}(\theta^\top x - y)^2$—is essentially optimal. Our goal in this section is thus to provide more understanding of potential poor behavior of the procedure (6) with

respect to ERM, considering two scenarios. The first is in stochastic (convex) optimization problems, where we investigate the finite-sample convergence rates of the robust solution to the population optimal risk. We show that the robust solution $\widehat{\theta}_n^{\rm rob}$ enjoys fast rates of convergence in cases in which the risk has substantial curvature—precisely as with empirical risk minimization. The second is to consider the asymptotics of the robust solution $\widehat{\theta}_n^{\rm rob}$, where we show that in classical statistical scenarios the robust solution is nearly efficient, though there is an asymptotic bias of order $1/\sqrt{n}$ that scales with the confidence ρ.

4.1. Fast Rates

In cases in which the risk R has curvature, empirical risk minimization often enjoys faster rates of convergence (Boucheron et al., 2005; Shapiro et al., 2009). The robust solution $\widehat{\theta}_n^{\rm rob}$ similarly attains faster rates of convergence in such cases, even with approximate minimizers of $R_n(\theta, \mathcal{P}_n)$. For the risk R and $\epsilon \geq 0$, let

$$S_\star^\epsilon := \left\{\theta \in \Theta : R(\theta) \leq \inf_{\theta^\star \in \Theta} R(\theta^\star) + \epsilon\right\}$$

denote the ϵ-sub-optimal (solution) set, and similarly let

$$\widehat{S}_\star^\epsilon := \left\{\theta \in \Theta : R_n(\theta, \mathcal{P}_n) \leq \inf_{\theta' \in \Theta} R_n(\theta', \mathcal{P}_n) + \epsilon\right\}.$$

For a vector $\theta \in \Theta$, let $\pi_{S_\star}(\theta) = \operatorname{argmin}_{\theta^\star \in S_\star} \|\theta^\star - \theta\|_2$ denote the Euclidean projection of θ onto the set $S_\star$; this projection operator is very useful for showing faster rates of convergence in stochastic optimization (see Shapiro et al. (2009), whose techniques we closely follow). In the statement of the result, for $A \subset \Theta$, we let $\mathfrak{R}_n(A)$ denote the Rademacher complexity of the localized process $\{x \mapsto \ell(\theta; x) - \ell(\pi_{S_\star}(\theta); x) : \theta \in A\}$. We then have the following result, whose proof we provide in Section E.

Theorem 8 *Let Θ be convex and let $\ell(\cdot; x)$ be convex and L-Lipshitz in its first argument for all $x \in \mathcal{X}$. For constants $\lambda > 0$, $\gamma > 1$, and $r > 0$, assume the risk R satisfies*

$$R(\theta) - \inf_{\theta \in \Theta} R(\theta) \geq \lambda \operatorname{dist}(\theta, S_\star)^\gamma \quad \text{for all } \theta \text{ such that } \operatorname{dist}(\theta, S_\star) \leq r. \tag{26}$$

Let $t > 0$. If $0 \leq \epsilon \leq \frac{1}{2}\lambda r^\gamma$ satisfies

$$\epsilon \geq \left(2\frac{8^\gamma L^\gamma}{\lambda}\right)^{\frac{1}{\gamma-1}} \left(\frac{\rho}{n}\right)^{\frac{\gamma}{2(\gamma-1)}} \quad \text{and} \quad \frac{\epsilon}{2} \geq 2\mathbb{E}[\mathfrak{R}_n(S_\star^{2\epsilon})] + L\left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}} \sqrt{\frac{2t}{n}}, \tag{27}$$

then $\mathbb{P}(\widehat{S}_\star^\epsilon \subset S_\star^{2\epsilon}) \geq 1 - e^{-t}$,

We provide a brief discussion of this result as well as a corollary that gives more explicit rates of convergence. First, we note that (by an inspection of the proof) the L-Lipschitz assumption need only hold in the neighborhood $S_\star^{2\epsilon}$ for the result to hold. We also have the following

Corollary 9 *In addition to the conditions of Theorem 8, assume that $S_\star = \{\theta^\star\}$ is a single point and $\Theta \subset \mathbb{R}^d$. Then for any $\epsilon \le \frac{1}{2}\lambda r^\gamma$, we have $\mathbb{P}(\widehat{S}_\star^\epsilon \subset S_\star^{2\epsilon}) \ge 1 - e^{-t}$ for*

$$\epsilon \gtrsim \left(\frac{L^\gamma}{\lambda}\right)^{\frac{1}{\gamma-1}} \left(\frac{d}{n}\log\frac{n}{d} + \frac{t}{n} + \frac{\rho}{n}\right)^{\frac{\gamma}{2(\gamma-1)}}.$$

So long as $\rho \lesssim d \log \frac{n}{d}$, this rate of convergence is as good as that enjoyed by standard empirical risk minimization approaches (Shapiro et al., 2009, Ch. 5) under these types of growth conditions. The case that $\gamma = 2$ corresponds (roughly) to strong convexity, and in this case we get the approximate rate of convergence of $\frac{L^2}{\lambda}\frac{d\log\frac{n}{d}}{n}$, the familiar rate of convergence under these conditions. Of course, if there is too much variance penalization (i.e. ρ is too large), then the rates of convergence may be slower.

Proof That $S_\star$ is a singleton implies that $S_\star^{2\epsilon} \subset \{\theta \mid \|\theta - \theta^\star\| \le (2\epsilon/\lambda)^{\frac{1}{\gamma}}\}$. Moreover, in this case we also have that

$$\left|\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X) - \ell(\theta^\star; X)] - \mathbb{E}_{\widehat{P}_n}[\ell(\theta'; X) - \ell(\theta^\star; X)]\right| \le L\left\|\theta - \theta'\right\|,$$

so that an ϵ/L-cover of $\{\theta \mid \|\theta - \theta^\star\| \le (2\epsilon/\lambda)^{\frac{1}{\gamma}}\}$ is an ϵ-cover of the function class $\mathcal{F} = \{f(x) = \ell(\theta; x) - \ell(\theta^\star; x) \mid \theta \in S_\star^{2\epsilon}\}$ in $\|\cdot\|_{L^2(P_n)}$ norm. Thus, the standard Dudley entropy integral (Dudley, 1999; van der Vaart and Wellner, 1996) yields

$$\begin{aligned}\mathbb{E}[\mathfrak{R}_n(S_\star^{2\epsilon})] &\lesssim \frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}, \delta, \|\cdot\|_{L^2(P_n)})}d\delta \\ &\lesssim \frac{1}{\sqrt{n}}\int_0^{L(2\epsilon/\lambda)^{\frac{1}{\gamma}}} \sqrt{d\log\frac{L}{\delta}}d\delta \le L\sqrt{\frac{d}{n}}\left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}}\sqrt{1 + \frac{1}{\gamma}\log\frac{\lambda}{2L^\gamma\epsilon}}\end{aligned}$$

where we have used that $\int_0^\varepsilon \sqrt{\log\frac{L}{\delta}}d\delta \le \varepsilon\sqrt{1 + \log\frac{L}{\varepsilon}}$. Solving for ϵ in the localization inequality (27) then yields the corollary, showing that the specified choice of ϵ is sufficient for all the conditions (27) to hold. ■

4.2. Asymptotics

It is important to understand the precise limiting behavior of the robust estimator in addition to its finite sample properties—this allows us to more precisely characterize when there may be *degradation* relative to classical risk minimization strategies. With that in mind, in this section we provide asymptotic results for the robust solution (6) to better understand the consequences of penalizing the variance of the loss itself. In particular, we would like to understand efficiency losses relative to (say) maximum likelihood in situations in which maximum likelihood is efficient. Before stating the results, we make a few standard assumptions on the risk $R(\theta)$, the loss ℓ, and the moments of ℓ and its derivatives. Concretely, we assume that

$$\theta^\star := \operatorname*{argmin}_\theta R(\theta) \quad \text{and} \quad \nabla^2 R(\theta^\star) \succ 0,$$

that is, the risk functional has strictly positive definite Hessian at $\theta^\star$, which is thus unique. Additionally, we have the following smoothness assumptions on the loss function, which are satisfied by common loss functions, including the negative log-likelihood for any exponential family or generalized linear model (Lehmann and Casella, 1998). In the assumption, we let $\mathbb{B}$ denote the ℓ_2-ball of radius 1 in $\mathbb{R}^d$.

Assumption A *For some $\epsilon > 0$, there exists a function $L : \mathcal{X} \to \mathbb{R}_+$ satisfying*

$$|\ell(\theta, x) - \ell(\theta', x)| \le L(x) \left\|\theta - \theta'\right\|_2 \quad \text{for } \theta, \theta' \in \theta^\star + \epsilon\mathbb{B}$$

and $\mathbb{E}[L(X)^2] \le L(P) < \infty$. Additionally, there is a function H such that the function $\theta \mapsto \ell(\theta, x)$ has $H(x)$-Lipschitz continuous Hessian (with respect to the Frobenius norm) on $\theta^\star + \epsilon\mathbb{B}$, where $\mathbb{E}[H(X)^2] < \infty$.

Then, recalling the robust estimator (6) as the minimizer of $R_n(\theta, \mathcal{P}_n)$, we have the following theorem, which we prove in Section F.

Theorem 10 *Let Assumption A hold, and let the sequence $\widehat{\theta}_n^{\rm rob}$ be defined by $\widehat{\theta}_n^{\rm rob} \in \operatorname{argmin}_\theta R_n(\theta, \mathcal{P}_n)$. Define*

$$b(\theta^\star) := \frac{\operatorname{Cov}(\nabla_\theta \ell(\theta^\star, X), \ell(\theta^\star, X))}{\sqrt{\operatorname{Var}(\ell(\theta^\star, X))}} \text{ and } \Sigma(\theta^\star) = \left(\nabla^2 R(\theta^\star)\right)^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star, X)) \left(\nabla^2 R(\theta^\star)\right)^{-1}.$$

Then $\widehat{\theta}_n^{\rm rob} \stackrel{a.s.}{\to} \theta^\star$ and

$$\sqrt{n}(\widehat{\theta}_n^{\rm rob} - \theta^\star) \stackrel{d}{\to} \mathsf{N}\left(-\sqrt{2\rho}\, b(\theta^\star), \Sigma(\theta^\star)\right)$$

The asymptotic variance $\Sigma(\theta^\star)$ in Theorem 10 is generally unimprovable, as made apparent by Le Cam's local asymptotic normality theory and the Hájek-Le Cam local minimax theorems (van der Vaart and Wellner, 1996). Thus, Theorem 10 shows that the robust regularized estimator (6) has some efficiency loss, but it is only in the bias term. We explore this a bit more in the context of the risk of $\widehat{\theta}_n^{\rm rob}$. Letting $W \sim \mathsf{N}(0, \Sigma(\theta^\star))$, as an immediate corollary to this theorem, the delta-method implies that

$$n\left[R(\widehat{\theta}_n^{\rm rob}) - R(\theta^\star)\right] \stackrel{d}{\to} \frac{1}{2}\left\|\sqrt{2\rho}\, b(\theta^\star) + W\right\|^2_{\nabla^2 R(\theta^\star)}, \tag{28}$$

where we recall that $\|x\|_A^2 = x^\top A x$. This follows from a Taylor expansion, because $\nabla R(\theta^\star) = 0$ and so $R(\theta) - R(\theta^\star) = \frac{1}{2}(\theta - \theta^\star)^\top \nabla^2 R(\theta^\star)(\theta - \theta^\star) + o(\|\theta - \theta^\star\|^2)$, or

$$\begin{aligned} n(R(\widehat{\theta}_n^{\rm rob}) - R(\theta^\star)) &= n\left(\frac{1}{2}(\widehat{\theta}_n^{\rm rob} - \theta^\star)^\top \nabla^2 R(\theta^\star)(\widehat{\theta}_n^{\rm rob} - \theta^\star) + o(\|\widehat{\theta}_n^{\rm rob} - \theta^\star\|^2)\right) \\ &= \frac{1}{2}\left(\sqrt{n}(\widehat{\theta}_n^{\rm rob} - \theta^\star)\right)^\top \nabla^2 R(\theta^\star)\left(\sqrt{n}(\widehat{\theta}_n^{\rm rob} - \theta^\star)\right) + o_P(1) \\ &\stackrel{d}{\to} \frac{1}{2}(\sqrt{2\rho}\, b(\theta^\star) + W)^\top \nabla^2 R(\theta^\star)(\sqrt{2\rho}\, b(\theta^\star) + W) \end{aligned}$$

by Theorem 10.

The limiting random variable in expression (28) has expectation

$$\frac{1}{2}\mathbb{E}[\|\sqrt{2\rho}b(\theta^\star) + W\|^2_{\nabla^2 R(\theta^\star)}] = \rho b(\theta^\star)^\top \nabla^2 R(\theta^\star) b(\theta^\star) + \frac{1}{2}\operatorname{tr}(\nabla^2 R(\theta^\star)^{-1} \operatorname{Cov}(\ell(\theta^\star, X)),$$

while the classical empirical risk minimization procedure (standard M-estimation) (Lehmann and Casella, 1998; van der Vaart and Wellner, 1996) has limiting mean-squared error $\frac{1}{2}\operatorname{tr}(\nabla^2 R(\theta^\star)^{-1} \operatorname{Cov}(\ell(\theta^\star, X)))$. Thus there is an additional $\rho \|b(\theta^\star)\|^2_{\nabla^2 R(\theta^\star)}$ penalty in the asymptotic risk (at a rate of $1/n$) for the robustly-regularized estimator. An inspection of the proof of Theorem 10 reveals that $b(\theta^\star) = \nabla_\theta \sqrt{\operatorname{Var}(\ell(\theta^\star, X))}$; if the variance of the loss is stable near $\theta^\star$, so that moving to a parameter $\theta = \theta^\star + \Delta$ for some small Δ has little effect on the variance, then the standard loss terms dominate, and robust regularization has asymptotically little effect. On the other hand, highly unstable loss functions for which $\nabla_\theta \sqrt{\operatorname{Var}(\ell(\theta^\star, X))}$ is large yield substantial bias.

We conclude our study of the asymptotics with a (to us) somewhat surprising example. Consider the classical linear regression setting in which $y = x^\top \theta^\star + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$. Using the standard squared error loss $\ell(\theta, (x, y)) = \frac{1}{2}(\theta^\top x - y)^2$, we obtain that

$$\nabla \ell(\theta^\star, (x, y)) = (x^\top \theta^\star - y)x = (x^\top \theta^\star - x^\top \theta^\star - \varepsilon)x = -\varepsilon x,$$

while $\ell(\theta^\star, (x, y)) = \frac{1}{2}\varepsilon^2$. The covariance $\operatorname{Cov}(\varepsilon X, \varepsilon^2) = \mathbb{E}[\varepsilon X(\varepsilon^2 - \sigma^2)] = 0$ by symmetry of the error distribution, and so—in the special classical case of correctly specified linear regression—the bias term $b(\theta^\star) = 0$ for linear regression in Theorem 10. That is, the robustly regularized estimator (6) is asymptotically efficient.

5. Experiments

We present three experiments in this section. The first is a small simulation example, which serves as a proof of concept allowing careful comparison of standard empirical risk minimization (ERM) strategies to our variance-regularized approach. The latter two are classification problems on real datasets; for both of these we compare performance of robust solution (6) to its ERM counterpart.

5.1. Minimizing the robust objective

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 (Hiriart-Urruty and Lemaréchal, 1993, Theorem VI.4.4.2) is that if the vector $p^* \in \mathbb{R}^n_+$

. Code is available at `https://github.com/hsnamkoong/robustopt`.

achieving the supremum in the definition (4) of the robust risk is unique, then

$$\partial_\theta R_n(\theta, \mathcal{P}_n) = \partial_\theta \sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(\theta; X)] = \sum_{i=1}^n p_i^* \partial_\theta \ell(\theta; X_i),$$

where the final summation is the standard Minkowski sum of sets. As this maximizing vector p is indeed unique whenever $\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X)) \neq 0$, we see that for all our problems, so long as ℓ is differentiable, so too is $R_n(\theta, \mathcal{P}_n)$ and

$$\nabla_\theta R_n(\theta, \mathcal{P}_n) = \sum_{i=1}^n p_i^* \nabla_\theta \ell(\theta; X_i) \quad \text{where} \quad p^* = \operatorname*{argmax}_{p \in \mathcal{P}_n} \left\{ \sum_{i=1}^n p_i \ell(\theta; X_i) \right\}. \tag{29}$$

In order to perform gradient descent on the risk $R_n(\theta, \mathcal{P}_n)$, then, by equation (29) we require only the computation of the worst-case distribution p^*. By taking the dual of the maximization (29), this is an efficiently solvable convex problem; for completeness, we provide a procedure for this computation in Section H that requires time $O(n \log n + \log \frac{1}{\epsilon} \log n)$ to compute an ϵ-accurate solution to the maximization (29). As all our examples have smooth objectives, we perform gradient descent on the robust risk $R_n(\cdot, \mathcal{P}_n)$, with stepsizes chosen by a backtracking (Armijo) line search (Boyd and Vandenberghe, 2004, Chapter 9.2).

5.2. Simulation experiment

For our simulation experiment, we use a quadratic loss with linear perturbation. For $v, x \in \mathbb{R}^d$, define the loss $\ell(\theta; x) = \frac{1}{2}\|\theta - v\|_2^2 + x^\top(\theta - v)$. We set $d = 50$ and take $X \sim \mathsf{Uni}(\{-B, B\}^d)$, varying B in the experiment. For concreteness, we let the domain $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$ and set $v = \frac{r}{2\sqrt{d}}\mathbf{1}$, so that $v \in \operatorname{int}\Theta$; we take $r = 10$. Notably, standard regularization strategies, such as ℓ_1 or ℓ_2-regularization, pull θ toward 0, while the variance of $\ell(\theta; X)$ is minimized by $\theta = v$ (thus naturally advantaging the variance-based regularization we consider, as $R(v) = \inf_\theta R(\theta) = 0$). Moreover, as X is pure noise, this is an example where we expect variance regularization to be particularly useful. We choose $\delta = .05$ and set ρ as in Eq. (17) (using that ℓ is $(3r + \sqrt{d}B)$-Lipschitz) to obtain robust coverage with probability at least $1 - \delta$. In our experiments, we obtained 100% coverage in the sense of (15), as the high probability bound is conservative.

Figure 2 summarizes the results. The robust solution $\widehat{\theta}_n^{\mathrm{rob}} = \operatorname{argmin}_{\theta \in \Theta} R_n(\theta, \mathcal{P}_n)$ always outperforms the empirical risk minimizer $\widehat{\theta}_n^{\mathrm{erm}} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)]$ in terms of the true risk $\mathbb{E}[\ell(\theta, X)] = \frac{1}{2}\|\theta - v\|_2^2$. Each experiment consists of 1,200 independent replications for each sample size n and value B. In Tables 1 and 2, we display the risks of $\widehat{\theta}_n^{\mathrm{erm}}$ and $\widehat{\theta}_n^{\mathrm{rob}}$ and variances, respectively, computed for the 1,200 independent trials. The gap between the risk of $\widehat{\theta}_n^{\mathrm{erm}}$ and $\widehat{\theta}_n^{\mathrm{rob}}$ is siginificant at level $p < .01$ for all sample sizes and values of B we considered according to a one-sided T-test. Notice also in Table 2 that the variance of the robust solutions is substantially smaller than that of the empirical risk minimizer—often several orders of magnitude smaller for large sample sizes n. This simulation shows that—in a simple setting favorable to it—our procedure outperforms standard alternatives.

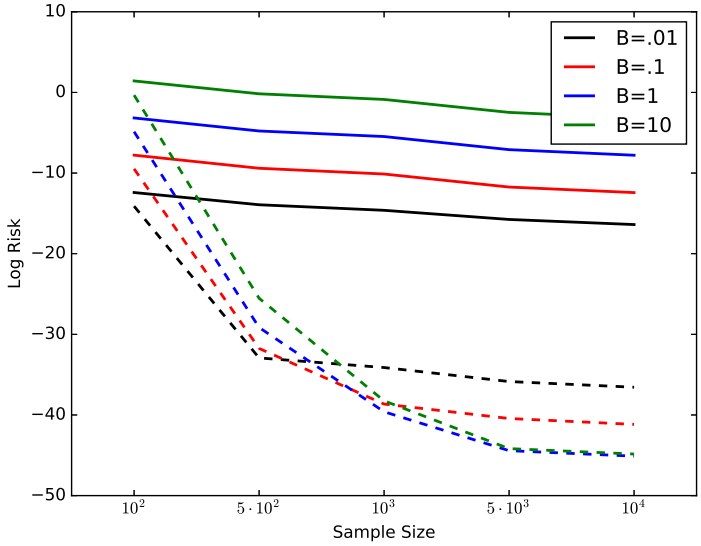


Figure 2. Simulation experiment. $\log \mathbb{E}[R(\widehat{\theta}_n^{\mathrm{erm}})]$ is the solid lines, in decreasing order from $B = 10$ (top) to $B = .01$ (bottom). $\log \mathbb{E}[R(\widehat{\theta}_n^{\mathrm{rob}})]$ is the dashed line, in the same vertical ordering at sample size $n = 10^2$.

Table 1: Simulation experiment: Mean risks over 1,200 simulations

	$B = .01$		$B = .1$		$B = 1$		$B = 10$	
n	$R(\widehat{\theta}_n^{\mathrm{erm}})$	$R(\widehat{\theta}_n^{\mathrm{rob}})$	$R(\widehat{\theta}_n^{\mathrm{erm}})$	$R(\widehat{\theta}_n^{\mathrm{rob}})$	$R(\widehat{\theta}_n^{\mathrm{erm}})$	$R(\widehat{\theta}_n^{\mathrm{rob}})$	$R(\widehat{\theta}_n^{\mathrm{erm}})$	$R(\widehat{\theta}_n^{\mathrm{rob}})$
100	4.06E-06	7.42E-07	4.17E-04	7.65E-05	4.20E-02	7.64E-03	4.15E+00	7.12E-01
500	8.91E-07	5.01E-15	8.22E-05	1.63E-14	8.36E-03	2.19E-13	8.41E-01	8.21E-12
1000	4.47E-07	1.52E-15	4.02E-05	1.64E-17	4.20E-03	6.32E-18	4.19E-01	2.45E-17
5000	1.44E-07	2.68E-16	8.00E-06	2.74E-18	8.27E-04	5.09E-20	8.38E-02	6.55E-20
10000	7.64E-08	1.32E-16	4.02E-06	1.32E-18	4.13E-04	2.57E-20	4.18E-02	3.34E-20

5.3. Protease cleavage experiments

For our second experiment, we compare our robust regularization procedure to other regularizers using the HIV-1 protease cleavage dataset from the UCI ML-repository (Lichman, 2013). 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 1360 examples with label $Y = +1$ (HIV-1 cleaves) and 5230 examples with $Y = -1$ (does not cleave).

We use the logistic loss $\ell(\theta; (x, y)) = \log(1+\exp(-y\theta^\top x))$. We compare the performance of different constraint sets Θ by taking

$$\Theta = \left\{\theta \in \mathbb{R}^d : a_1 \left\|\theta\right\|_1 + a_2 \left\|\theta\right\|_2 \leq r\right\},$$

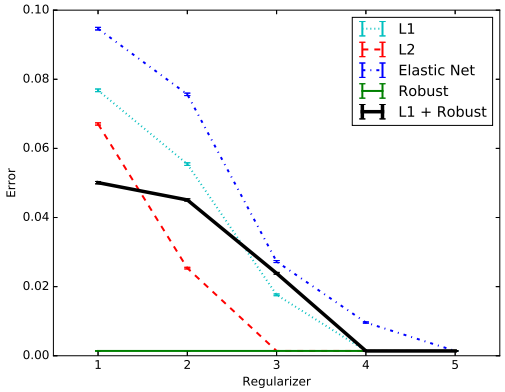
Table 2: Simulation experiment: Variances of $R(\widehat{\theta})$ over 1,200 simulations

	$B = .01$		$B = .1$		$B = 1$		$B = 10$	
n	ERM	Robust	ERM	Robust	ERM	Robust	ERM	Robust
100	7.06E-13	9.76E-14	6.58E-09	1.03E-09	7.09E-05	1.08E-05	7.37E-01	9.20E-02
500	5.98E-14	7.15E-28	3.04E-10	3.52E-26	2.80E-06	2.26E-24	2.92E-02	3.26E-21
1000	2.63E-14	1.07E-31	7.53E-11	1.99E-35	7.14E-07	3.44E-33	7.03E-03	4.78E-32
5000	7.34E-15	2.94E-33	2.70E-12	3.28E-37	2.95E-08	2.50E-39	2.74E-04	5.24E-38
10000	1.60E-15	6.54E-34	6.74E-13	7.59E-38	7.04E-09	3.34E-39	6.52E-05	2.25E-38

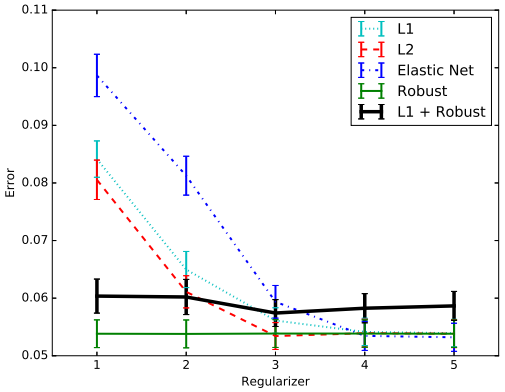
which is equivalent to elastic net regularization (Zou and Hastie, 2005), while varying a_1, a_2, and r. We experiment with ℓ_1-constraints ($a_1 = 1, a_2 = 0$) with $r \in \{50, 100, 500, 1000, 5000\}$, ℓ_2-constraints ($a_1 = 0, a_2 = 1$) with $r \in \{5, 10, 50, 100, 500\}$, elastic net ($a_1 = 1, a_2 = 10$) with $r \in \{100, 200, 1000, 2000, 10000\}$, our robust regularizer with $\rho \in \{100, 1000, 10000, 50000, 100000\}$ and our robust regularizer coupled with the ℓ_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 0-1 (misclassification) loss $1\{\text{sign}(\theta^T x)y \le 0\}$. 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 fraction (test).

We plot results summarizing these experiments in Figure 3. The horizontal axis in each figure indexes our choice of regularization value (so "`Regularizer = 1`" for the ℓ_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 $\widehat{\theta}_n^{\text{rob}}$ achieves low classification error for all values of ρ (Figure 3(b)). Notably, even when coupled with a fairly stringent ℓ_1-constraint ($r = 100$), robust regularization has perofrmance better than ℓ_1 except for large values r, especially on the rare label $Y = +1$ (Figure 3 (d) and (f)).

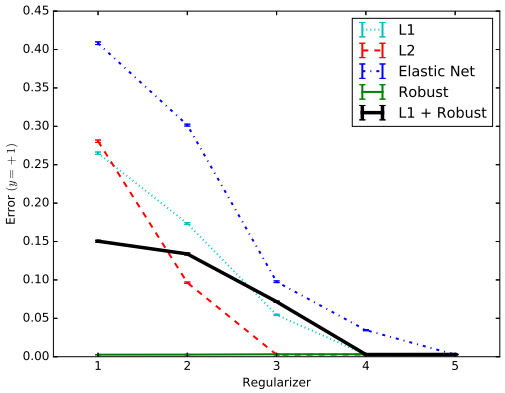
We investigate the effects of the robust regularizer with a slightly different perspective in Figure 4, where we use $\Theta = \{\theta : \|\theta\|_1 \le r\}$ with $r = 100$ for the constraint set for each experiment. The horizontal axis indicates the tolerance ρ we use in construction of the robust estimator $\widehat{\theta}_n^{\text{rob}}$, where `ERM` means $\rho = 0$. In Fig. 4(a), we plot the logistic risk $R(\widehat{\theta}) = \mathbb{E}[\ell(\widehat{\theta}, (X, Y))]$ for the train and test distribution. We also plot the upper confidence bound $R_n(\theta, \mathcal{P}_n)$ in this plot, which certainly over-estimates the test risk—we hope to tighten this overestimate in future work. In Figure 4(b), we plot the misclassification error on train and test for different values of ρ, along with 2-standard-error intervals for the 50 runs. Figures 4(c) and (d) show the error rates restricted to examples from the uncommon (c) and common (d) classes. In Table 3 we give explicit error rates and logistic risk values for the different procedures. Due to the small size of the test dataset ($n_{\text{test}} = 659$), the deviation across folds is somewhat large.



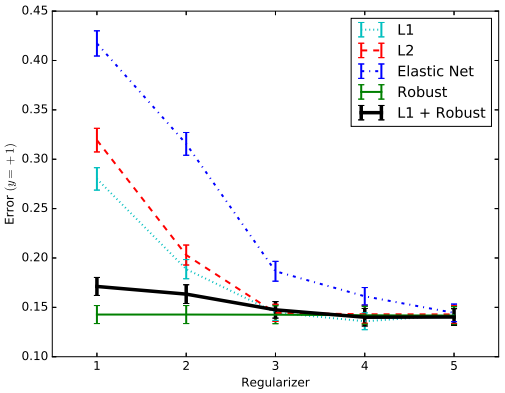
(a) Train error



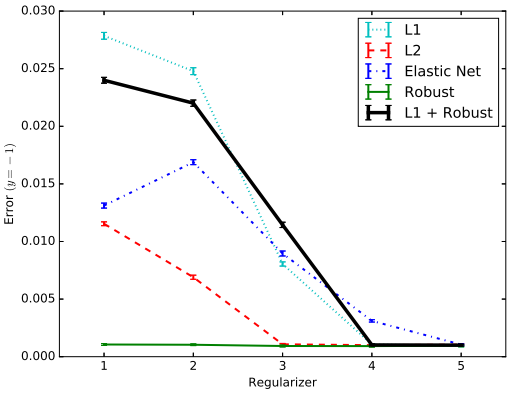
(b) Test error



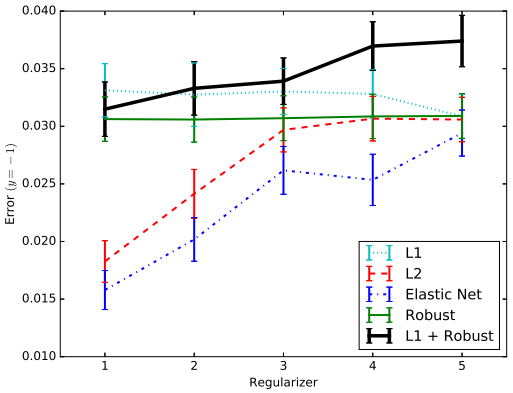
(c) Train error on rare class ($Y_i = +1$)



(d) Test error on rare class ($Y_i = +1$)

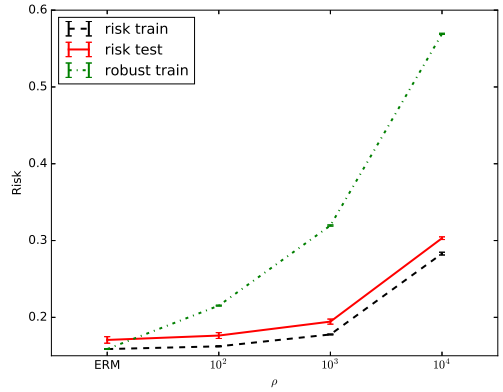


(e) Train error on common class ($Y_i = -1$)

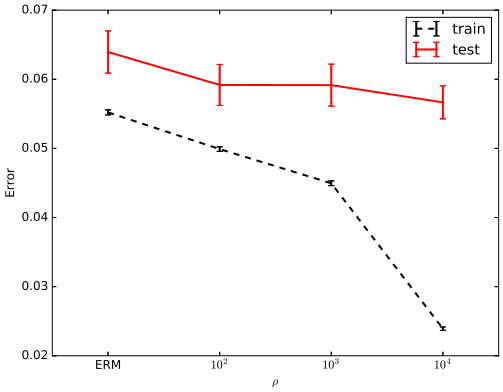


(f) Test error on common class ($Y_i = -1$)

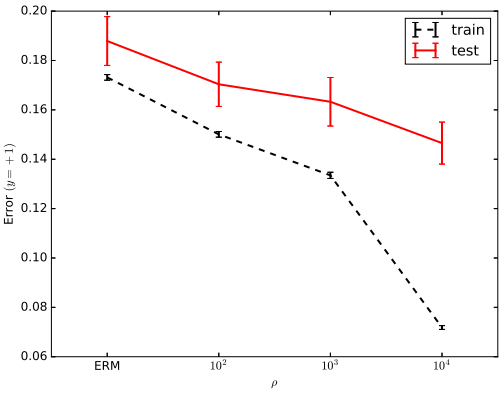
Figure 3. HIV-1 Protease Cleavage plots (2-standard error confidence bars). Comparison of misclassification error rates among different regularizers.



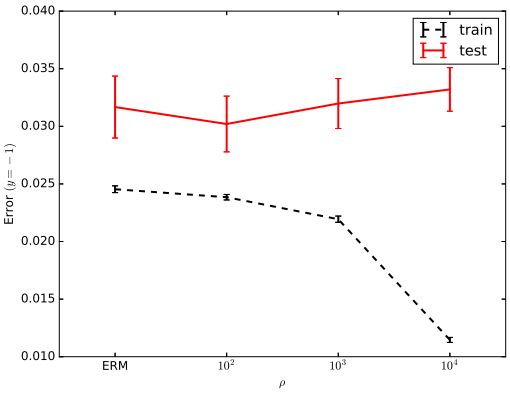
(a) Logistic risk and confidence bound



(b) Misclassification error rate



(c) Error on rare class ($Y_i = +1$)



(d) Error on common class ($Y_i = -1$)

Figure 4. HIV-1 Protease Cleavage plots (2-standard error confidence bars). Plot (a) shows the logistic risk $R(\theta) = \mathbb{E}[\log(1+e^{-Y\theta^\top X})]$ and confidence bounds computed from the robust risk (4). Plots (b)–(d) show misclassification error rates plotted against robustness parameter ρ.

In this experiment, we see (roughly) that the ERM solutions achieve good performance on the common class ($Y = -1$) but sacrifice performance on the uncommon class. As we increase ρ, performance of the robust solution $\widehat{\theta}_n^{\rm rob}$ on the rarer label $Y = +1$ improves (Fig. 4(c)), while the misclassification rate on the common class degrades a small (insignificant) amount (Fig. 4(d)); see also Table 3. This behavior is roughly what we might expect for the robust estimator: the poor performance of the ERM estimator $\widehat{\theta}_n^{\rm erm}$ on the rare class induces (relatively) more variance, which the robust solution reduces by via improved classification performance on the rare ($Y = +1$) class. This occurs at little expense over the more common label $Y = -1$ so that overall performance improves by a small amount. We remark—but are unable to explain—that this improvement on classification error for the rare labels comes despite increases in logistic risk; while the average logistic loss increases, misclassification errors decrease.

Table 3: HIV-1 Cleavage Error

	risk		error (%)		error ($Y = +1$)		error ($Y = -1$)	
ρ	train	test	train	test	train	test	train	test
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

5.4. Document classification in the Reuters corpus

For our final experiment, we consider a multi-label classification problem with a reasonably large dataset. The Reuters RCV1 Corpus (Lewis et al., 2004) 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 Corporate, Economics, Goverment, and Markets, and each document in the data is labeled with a subset of those. As each document can belong to multiple categories, we fit binary classifiers on each of the four categories. There are different numbers of documents labeled as each category, with the Economics category having the fewest number of positive examples. Table 4 gives the number of times a document is labeled as each of the four categories (so each document has about 1.18 associated classes). In this experiment, we expect the robust solution to outperform ERM on the rarer category (Economics), as the robustification (6) naturally upweights rarer (harder) instances, which disproportionally affect variance—as in the experiment on HIV-1 cleavage.

Table 4: Reuters Number of Examples

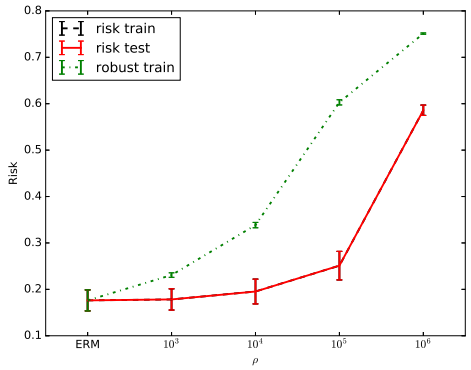
Corporate	Economics	Government	Markets
381,327	119,920	239,267	204,820

For each category $k \in \{1, 2, 3, 4\}$, we use the logistic loss $\ell(\theta_k; (x, y)) = \log(1+\exp(-y\theta_k^\top x))$. For each binary classifier, we use the ℓ_1 constraint set $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \le 1000\right\}$. 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):

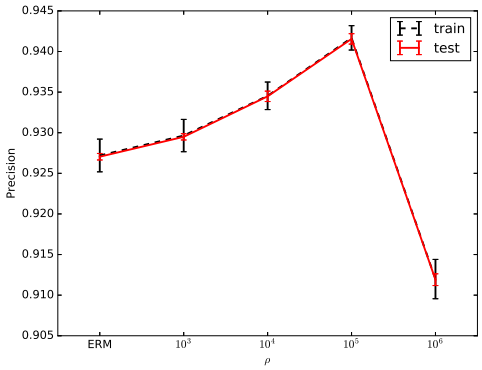
$$\text{precision} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=1}^{4} 1\{\theta_k^\top x_i \ge 0, y_i = 1\}}{\sum_{k=1}^{4} 1\{\theta_k^\top x_i > 0\}},$$
$$\text{recall} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=1}^{4} 1\{\theta_k^\top x_i \ge 0, y_i = 1\}}{\sum_{k=1}^{4} 1\{y_i = 1\}}.$$

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.

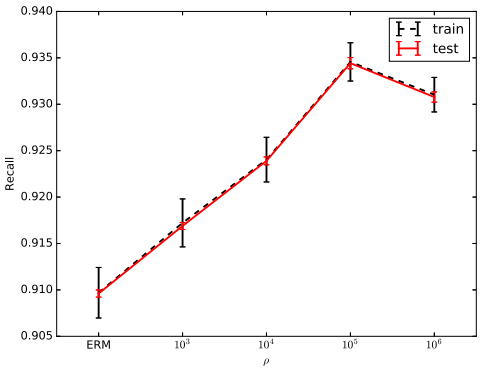
In Figure 5, we summarize the results of our experiment averaged over the 10 runs, with 2-standard error bars (computed across the folds). To facilitate comparison across the document categories, we give exact values of these averages in Tables 5 and 6. Both



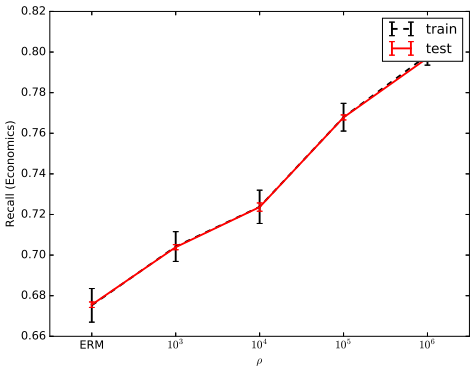
(a) Logistic risk and confidence bound



(b) Precision



(c) Recall



(d) Recall on rare category (Economics)

Figure 5: Reuters Corpus (2-standard error deviations)

$\widehat{\theta}_n^{\rm rob}$ and $\widehat{\theta}_n^{\rm erm}$ have reasonably high precision across all categories, with increasing ρ giving a mild improvement in precision (from $.93 \pm .005$ to $.94 \pm .005$); see also Figure 5(a). On the other hand, we observe in Figure 5(d) that ERM has low recall (.69 on test) for the Economics category, which contains about 15% of documents. As we increase ρ from 0 (ERM) to 10^5, 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. We also note that while the robust solutions outperform ERM in classification performance for $\rho \le 10^5$, for very large $\rho = 10^6 \ge 10n$, the regularizing effects of robustness degrade the solution $\widehat{\theta}_n^{\rm rob}$. This precision and recall improvement comes in spite of the increase in the average binary logistic risk for each of the 4 classes, which we show in Figure 5a, which plots the average binary logistic loss (on train and test sets) averaged over the 4 categories as well as the upper confidence bound $R_n(\theta, \mathcal{P}_n)$ as we vary ρ. The robust regularization effects reducing variance appear to improve the performance of the binary logistic loss as a surrogate for true misclassification error.

Table 5: Reuters Corpus Precision (%)

	Precision		Corporate		Economics		Government		Markets	
ρ	train	test	train	test	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 6: Reuters Corpus Recall (%)

	Recall		Corporate		Economics		Government		Markets	
ρ	train	test	train	test	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

Summary

We have seen through multiple examples that robustification—our convex surrogate for variance regularization—is an effective tool in a number of applications. As we heuristically expect, variance-based regularization (robust regularization) yields predictors with better performance on "hard" instances, or subsets of the problem that induce higher variance, such as classes with relatively few training examples in classification problems. The robust regularization ρ gives a principled knob for tuning performance to trade between variance (uniform or across-the-board performance) and—sometimes—absolute performance.

6. Discussion

In this paper, we have developed theoretical results for robust regularization (6) that apply to general stochastic optimization and learning problems problems. The examples we describe in Section 3 illustrate our expectation that the robust solution $\widehat{\theta}_n^{\rm rob}$ should have good performance in cases in which $\mathrm{Var}(\ell(\theta^\star; X))$ is small (recall also Theorems 3 and 6). Identifying the separation between the performance empirical risk minimization and related estimators and that of the robustly-regularized estimators—as well as variance-regularized estimates—we consider more generally remains a challenge. We hope that this paper inspires work in this direction in machine learning and statistics, and more broadly, torward considering distributionally robust problems. Part of this is likely to come from making rigorous our empirical observations (Section 5) that robust regularization improves performance on "hard" instances without sacrificing performance on easier cases.

Our understanding of so-called "fast rates" for stochastic optimization problems, while considering robustness, is also limited. For empirical risk minimization, fast rates of convergence hold under conditions in which the the gap $R(\theta) - R(\theta^\star)$ controls the variance of the excess loss $\ell(\theta, X) - \ell(\theta^\star, X)$ (cf. Mammen and Tsybakov, 1999; Bartlett et al., 2005;

Boucheron et al., 2005; Bartlett et al., 2006), which usually requires some type of uniform convexity assumption. These bounds typically follow from localization guarantees (Bartlett et al., 2005, Section 5) on the function class

$$\{x \mapsto \ell(\theta, x) - \ell(\theta^\star, x) \mid \theta \in \Theta\}.$$

While in Section 4.1, we show that the robust estimate $\widehat{\theta}_n^{\rm rob}$ enjoys faster rates of convergence under growth conditions analogous to uniform convexity of the risk, as $\mathrm{Var}(\ell(\theta; X) - \ell(\theta^\star; X)) \neq \mathrm{Var}(\ell(\theta; X))$, it is not clear how to directly connect these guarantees to results of the form in Theorems 3 and 6. We leave investigation of these topics to future work.

The last point of our discussion is to revisit Theorem 6, which provides a guarantee for robustly regularized estimators based on localized Rademacher complexities. An investigation of our proof shows that our derivation proceeds by considering the complexity of self-normalized classes of functions of the form

$$\mathcal{G}_r = \left\{ \sqrt{\frac{r}{\mathbb{E}[f^2] \vee r}} f \mid f \in \mathcal{F} \right\}.$$

In contrast, the analogous result of Bartlett et al. (2005, Thereom 3.3) for empirical risk minimization considers the complexity of classes of functions of the form

$$\mathcal{G}_r = \left\{ \frac{r}{\mathbb{E}[f^2] \vee r} f \mid f \in \mathcal{F} \right\}.$$

The latter class normalizes functions f by $\sqrt{\mathbb{E}[f^2]}$—a type of self-normalization that arises in the computation of pivotal (asymptotically independent of the underlying distribution) statistics. While this choice *prima facie* is just a step in our proof, the robust objective $R_n(\theta, \mathcal{P}_n)$ defined in Eq. (4) is an empirical likelihood upper confidence bound on the optimal population risk (see also Duchi et al., 2016). One of the important characteristics of empirical likelihood confidence bounds is that they are self-normalizing and yield pivotal statistics (Owen, 2001). Investigating such self-normalization in complexity guarantees seems likely to yield fruitful insights.

Acknowledgments

We thank Feng Ruan for pointing out a much simpler proof of Theorem 1 than in our original paper. JCD and HN were partially supported by the SAIL-Toyota Center for AI Research and HN was partially supported by the Samsung Fellowship. JCD was also partially supported by the National Science Foundation award NSF-CAREER-1553086 and the Sloan Fellowship.

Appendix A. Proof of Theorem 1

The theorem is immediate if $s_n = 0$ or $\sigma^2 = 0$, as in this case $\sup_{P: D_\phi(P\|\widehat{P}_n) \le \rho/n} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] = \mathbb{E}[Z]$. In what follows, we will thus assume that $\sigma^2, s_n^2 > 0$. We recall the maximization problem (8), which is

$$\underset{p}{\text{maximize}} \ \sum_{i=1}^n p_i z_i \ \ \text{subject to} \ \ p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}_+^n : \frac{1}{2} \left\| np - \mathbf{1} \right\|_2^2 \le \rho, \langle \mathbf{1}, p \rangle = 1 \right\},$$

and the solution criterion (9), which guarantees that the maximizing value of problem (8) is $\overline{z} + \sqrt{2\rho s_n^2/n}$ whenever

$$\sqrt{2\rho} \frac{z_i - \overline{z}}{\sqrt{n s_n^2}} \ge -1.$$

Letting $z = Z$, then under the conditions of the theorem, we have $|z_i - \overline{z}| \le M$, and to satisfy inequality (9) it is certainly sufficient that

$$2\rho \frac{M^2}{n s_n^2} \le 1, \ \ \text{or} \ \ n \ge \frac{2\rho M^2}{s_n^2}, \ \ \text{or} \ \ s_n^2 \ge \frac{2\rho M^2}{n}. \tag{30}$$

Conversely, suppose that $s_n^2 < \frac{2\rho M^2}{n}$. Then we have $\frac{2\rho s_n^2}{n} < \frac{4\rho^2 M^2}{n^2}$, which in turn implies that

$$\sup_{p \in \mathcal{P}_n} \langle p, z \rangle \ge \frac{1}{n} \langle \mathbf{1}, z \rangle + \left(\sqrt{\frac{2\rho s_n^2}{n}} - \frac{2M\rho}{n} \right)_+ .$$

Combining this inequality with the condition (30) for the exact expansion to hold yields the two-sided variance bounds (10).

We now turn to showing the high-probability exact expansion (11), which occurs whenever the sample variance is large enough by expression (30). To that end, we show that s_n^2 is bounded from below with high probability. Define the event

$$\mathcal{E}_n := \left\{ s_n^2 \ge \frac{3}{64} \sigma^2 \right\},$$

and let $n \ge \frac{4M^2}{\sigma^2} \max\{2\sigma, 11\}$. Then, on event $\mathcal{E}_n$ we have $n \ge \frac{44\rho M^2}{\sigma^2} \ge \frac{2\rho M^2}{s_n^2}$, so that the sufficient condition (30) holds and expression (11) follows. We now argue that the event $\mathcal{E}_n$ has high probability via the following lemma due to Maurer and Pontil.

Lemma 11 (Maurer and Pontil (2009, Theorem 10)) *Let Z_i be i.i.d. random variables taking values in $[0, M]$, and let $s_n^2 = \frac{1}{n} \sum_{i=1}^n Z_i^2 - \left(\frac{1}{n} \sum_{i=1}^n Z_i \right)^2$. Then, for $n \ge 2$*

$$\mathbb{P}\left(s_n \le \sigma - t \right) \vee \mathbb{P}\left(s_n \ge \sigma + t \right) \le \exp\left(-\frac{n t^2}{2M^2} \right).$$

Setting $t = \left(1 - \frac{\sqrt{3}}{8}\right) \sigma$, the final result follows from noting that $\mathbb{P}(\mathcal{E}_n) \ge 1 - \exp\left(-\frac{nt^2}{2M^2}\right)$ from the lemma.

Appendix B. Proof of Theorem 2

Our starting point is to recall from inequality (30) in the proof of Theorem 1 that for each $f \in \mathcal{F}$, the empirical variance equality (12) holds if $n \geq \frac{4\rho M^2}{\mathrm{Var}_{\widehat{P}_n}(f)}$. As a consequence, Theorem 2 will follow if we can provide a uniform lower bound on the sample variances $\mathrm{Var}_{\widehat{P}_n}(f)$ that holds with high enough probability. We use C to denote a universal constant whose value may change from line to line. Noting that $\mathrm{Var}_{\widehat{P}_n}(f) = \mathbb{E}_{\widehat{P}_n}(f-\mathbb{E}[f])^2 - (\mathbb{E}_{\widehat{P}_n}(f-\mathbb{E}[f]))^2$, we proceed in two parts. First, we give a lower bound for $\mathbb{E}_{\widehat{P}_n}(f-\mathbb{E}[f])^2$.

Lemma 12 *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [M_0, M_1]$ with $M := M_1 - M_0$. Then, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$*

$$\mathrm{Var}(f) \leq 2\mathbb{E}_{\widehat{P}_n}(f - \mathbb{E}[f])^2 + C\left[\mathfrak{R}_n^{\sup}(\mathcal{F})^2 \log^3(nM) + \frac{M^2}{n}\left(t + \log\log n\right)\right].$$

Proof We follow the arguments of Srebro et al. (2010) and Bousquet (2002b, Thm. 6.1). For $x_1, \ldots, x_n \in \mathcal{X}$, let

$$\mathcal{F}_{n,r} := \left\{f - \mathbb{E}[f] \mid f \in \mathcal{F}, \;\; \mathbb{E}_{\widehat{P}_n}[(f - \mathbb{E}[f])^2] \leq r\right\},$$

where $\widehat{P}_n$ is the empirical measure on $x_1, \ldots, x_n$. Let $\psi_n^{\sup}$ be a sub-root upper bound on the worst-case Rademacher complexity

$$\psi_n^{\sup}(r) \geq \mathfrak{R}_n^{\sup}(\mathcal{F}_{n,r}),$$

where implicitly in the right hand side we take the supremum over $x_1, \ldots, x_n$ defininig $\mathcal{F}_{n,r}$ as well.

Lemma 13 (Srebro et al. (2010, Lemma 2.2)) *Let $\mathcal{H}$ be a class of bounded functions $\mathcal{X} \to [-M, M]$, and let $\kappa : [-M, M] \to \mathbb{R}_+$ be a bounded function with L-Lipschitz derivatives. Then,*

$$\mathfrak{R}_n^{\sup}\left(\left\{\kappa \circ h : h \in \mathcal{H}, \mathbb{E}_{\widehat{P}_n}[\kappa \circ h] \leq r\right\}\right) \leq C\sqrt{r}\mathfrak{R}_n^{\sup}(\mathcal{H}) \log^{\frac{3}{2}}(Mn).$$

Since $\kappa(t) = t^2$ has Lipschitz derivatives, above lemma with $\mathcal{H} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$ yields

$$\mathfrak{R}_n^{\sup}(\mathcal{F}_{n,r}^2) \leq C\sqrt{r}\mathfrak{R}_n^{\sup}(\mathcal{F}) \log^{\frac{3}{2}}(nM) \tag{31}$$

where we recall the notation that $\mathcal{G}^2 = \{g^2 \mid g \in \mathcal{G}\}$ for any function class $\mathcal{G}$. Thus we may take $\psi_n^{\sup}(r) = C\sqrt{r}\mathfrak{R}_n^{\sup}(\mathcal{F}) \log^{\frac{3}{2}} n$, which has fixed point $r_n^{\sup} = C^2\mathfrak{R}_n^{\sup}(\mathcal{F})^2 \log^3 n$. The following classical result then shows that the fixed point $r_n^{\sup}$ controls generalization of class of functions $\mathcal{F}_{n,r}^2$. Since $f^2 \geq 0$, Theorem 6.1 of Bousquet (2002b) yields that for all $f \in \mathcal{F}$,

$$\mathbb{E}(f - \mathbb{E}[f])^2 \leq 2\mathbb{E}_{\widehat{P}_n}(f - \mathbb{E}[f])^2 + C\left[\mathfrak{R}_n^{\sup}(\mathcal{F})^2 \log^3 nM + \frac{M^2}{n}\left(t + \log\log n\right)\right]$$

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

Next, we give an upper bound for $(\mathbb{E}_{\widehat{P}_n}(f - \mathbb{E}[f]))^2$. We use the following version of Talagrand's inequality due to Bousquet (2002a, 2003). (See also Bartlett et al. (2005, Thm 2.1).)

Lemma 14 *Let $r > 0$ and $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[a, b]$ such that for every $f \in \mathcal{F}$, $\mathrm{Var}(f(X)) \le r$. Then, with probability at least $1 - e^{-t}$*

$$\sup_{f\in\mathcal{F}}\{\mathbb{E}[f] - \mathbb{E}_{\widehat{P}_n}[f]\} \le \inf_{\alpha>0}\left\{2(1+\alpha)\mathbb{E}[\mathfrak{R}_n(\mathcal{F})] + \sqrt{\frac{2rt}{n}} + \frac{t}{n}(b-a)\left(\frac{1}{3}+\frac{1}{\alpha}\right)\right\}.$$

The same statement holds with $\sup_{f\in\mathcal{F}}(\mathbb{E}_{\widehat{P}_n}[f] - \mathbb{E}[f])$ replacing the left-hand side of the inequalities.

Applying Lemma 14 and letting $\alpha = \frac{1}{2}$, with probability at least $1 - 2e^{-t}$

$$|\mathbb{E}_{\widehat{P}_n}[f] - \mathbb{E}[f]| \le 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F})] + 2M\sqrt{\frac{2t}{n}}$$

holds for all $f \in \mathcal{F}$. Combining the above display with Lemma 12, we obtain the desired result.

Appendix C. Proof of Theorem 3

Before proving the theorem proper, we state a technical lemma that provides uniform Bernstein-like bounds for the class $\mathcal{F}$ using empirical ℓ_∞-covering numbers.

Lemma 15 (Maurer and Pontil (2009, Theorem 6)) *Let $n \ge \frac{8M^2}{t}$ and $t \ge \log 12$. Then with probability at least $1 - 6N_\infty(\mathcal{F}, \epsilon, 2n)e^{-t}$, we have*

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + 3\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f)t}{n}} + \frac{15Mt}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \tag{32}$$

for all $f \in \mathcal{F}$.

We return to the proof of Theorem 3. Let $\mathcal{E}_1$ denote that the event that the inequalities (32) hold. Then on $\mathcal{E}_1$ hold, uniformly over $f \in \mathcal{F}$ we have

$$\begin{aligned}
\mathbb{E}[f] &\le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{18\mathrm{Var}_{\widehat{P}_n}(f(X))t}{n}} + \frac{15Mt}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \\
&\overset{(i)}{\le} \sup_{P:D_\phi(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[f(X)] + \sqrt{\frac{2\rho\mathrm{Var}_{\widehat{P}_n}(f(X))}{n}} \\
&\qquad - \left(\sqrt{\frac{2\rho\mathrm{Var}_{\widehat{P}_n}(f(X))}{n}} - \frac{2M\rho}{n}\right)_+ + \frac{5M\rho}{3n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \\
&\le \sup_{P:D_\phi(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[f(X)] + \frac{11}{3}\frac{M\rho}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \text{ for all } f \in \mathcal{F},
\end{aligned} \tag{33}$$

where inequality (i) follows from the bounds (10) in Theorem 1 and the fact that $\rho \ge 9t$ by assumption. This gives the first result (15).

For the second result (16), we recall that $\widehat{f} \in \operatorname{argmin}_{f\in\mathcal{F}} \sup_P\{\mathbb{E}_P[f(X)] : D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}\}$, and we bound the supremum term in expression (33). First, we note that because $\widehat{f}$ minimizes the supremum term in expression (33), we have

$$\mathbb{E}[\widehat{f}] \le \sup_{P:D_\phi(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[f(X)] + \frac{11M\rho}{3n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \text{ for all } f \in \mathcal{F}.$$

Now fix $f \in \mathcal{F}$. As the function f is fixed, by Bernstein's inequality, we have

$$\mathbb{E}_{\widehat{P}_n}[f] \le \mathbb{E}[f] + \sqrt{\frac{2\mathrm{Var}(f)t}{n}} + \frac{2Mt}{3n}$$

with probability at least $1 - e^{-t}$. Similarly, we have by Lemma 11 that

$$\sqrt{\mathrm{Var}_{\widehat{P}_n}(f)} \le \sqrt{\mathrm{Var}(f)} + \sqrt{\frac{2tM^2}{n}}$$

with probability at least $1 - e^{-t}$. That is, for any fixed $f \in \mathcal{F}$, we have with probability at least $1 - 2e^{-t}$ that

$$\begin{aligned}
\sup_{P:D_\phi(P\|\widehat{P}_n)\le\frac{\rho}{n}} \mathbb{E}_P[f(X)] &\overset{(i)}{\le} \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\rho\mathrm{Var}_{\widehat{P}_n}(f)}{n}} \\
&\le \mathbb{E}[f] + \sqrt{\frac{2\mathrm{Var}(f)t}{n}} + \frac{2M}{3n}t + \sqrt{\frac{2\rho\mathrm{Var}(f)}{n}} + \frac{2\sqrt{M^2\rho t}}{n} \\
&\overset{(ii)}{\le} \mathbb{E}[f] + 2\sqrt{\frac{2\mathrm{Var}(f)\rho}{n}} + \frac{8}{3}\frac{M\rho}{n},
\end{aligned}$$

where inequality (i) follows from the uniform upper bound (10) of Theorem 1 and inequality (ii) from our assumption that $\rho \ge t$. Substituting this expression into our earlier bound (33) yields that for any $f \in \mathcal{F}$, with probability at least

$$1 - 2(3N_\infty(\mathcal{F}, \epsilon, 2n) + 1)e^{-t},$$

we have

$$\mathbb{E}[\widehat{f}(X)] \le \mathbb{E}[f(X)] + 2\sqrt{\frac{2\rho\mathrm{Var}(f(X))}{n}} + \frac{19}{3}\frac{M\rho}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon.$$

This gives the theorem.

Appendix D. Proof of Theorem 6

We first show the following version of uniform Bernstein's inequality with Rademacher complexities. The proof uses a peeling technique (Bartlett et al., 2005; van de Geer, 2000), in conjuction with Talagrand's concentration inequality (Lemma 14).

Lemma 16 *Let $r > 0$ and $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [0, M]$ with $\mathrm{Var}(f(X)) \le r$. Then, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$*

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2e\mathrm{Var}(f)}{n}\left(t + \log\left\lceil \log \frac{nr}{M^2 t}\right\rceil\right)} + 6\mathbb{E}[\mathfrak{R}_n(\mathcal{F})] + \frac{7M}{n}\left(t + \log\left\lceil \log \frac{nr}{M^2 t}\right\rceil\right).$$

The same statements hold with the roles of $\mathbb{E}[f]$ and $\mathbb{E}_{\widehat{P}_n}[f]$ reversed.

We defer the proof to section D at the end of this section. Because $\mathrm{Var}(f) \le M^2$ for all $f \in \mathcal{F}$, Lemma 16 also holds if we replace the terms $\left\lceil \log \frac{nr}{M^2 t}\right\rceil$ with $\left\lceil \log \frac{n}{t}\right\rceil \le 1 + \log \frac{n}{t}$.

Next, we show an important extension of Lemma 16 that replaces the Rademacher complexity term $\mathbb{E}[\mathfrak{R}_n(\mathcal{F})]$ by a local quantity $r_n^\star$, the fixed point of $\psi_n(r)$. To this end, we use another peeling argument and apply Lemma 16 to the self-normalized class

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

This idea follows the techniques of Bartlett et al. (2005, Thm. 3.3), though we use a type of self-normalizing scale, that is, $f/\sqrt{\mathbb{E}[f^2]}$, whereas they use a variance-normalizing scaling by studying classes of functions of the form $f/\mathbb{E}[f^2]$. Our use of this alternative normalization is important in the next lemma, which allows us to obtain bounds that apply to the robustly regularized risk.

Lemma 17 *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [0, M]$ satisfying the localization inequality (20) for some sub-root function $\psi_n(\cdot)$ with root $r_n^\star$. Let $B_n = \frac{1}{n}\left(t + \log\left\lceil \log \frac{n}{t}\right\rceil\right)$. Then with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$*

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \left(\sqrt{2eB_n} + 6\sqrt{r_n^\star + 7MB_n/3}\right)\sqrt{\mathbb{E}[f^2]} + 6r_n^\star + 14MB_n.$$

The same statement holds with the roles of $\mathbb{E}[f]$ and $\mathbb{E}_{\widehat{P}_n}[f]$ reversed.

See Section D.1 for the proof.

Next, we give an analogous result for f^2.

Lemma 18 *Let $\mathcal{F}$ be a collection of bounded functions $f : \mathcal{X} \to [0, M]$ satisfying the localization inequality (20) for some sub-root function $\psi_n(\cdot)$ with root $r_n^\star$. Let $\eta > 0$. Then, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$*

$$\mathbb{E}[f^2] \le \mathbb{E}_{\widehat{P}_n}[f^2] + \frac{1}{\eta}\mathbb{E}_{\widehat{P}_n}[f^2] + 72M^2(1+\eta)r_n^\star + \frac{Mt}{n}\left(4 + \frac{7}{3}M\right).$$

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

$$\mathbb{E}_{\widehat{P}_n}[f^2] \le \mathbb{E}[f^2] + \frac{\eta}{1+\eta}\mathbb{E}[f^2] + 72M^2(1+\eta)r_n^\star + \frac{Mt}{n}\left(4 + \frac{7}{3}M\right).$$

See Section D.2 for the proof.

Now, we make two additional pieces of shorthand notation. Let

$$V_n = 4((2e + 84M)B_n + 36r_n^\star).$$

Then, Lemma 17 implies that

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{V_n \mathbb{E}[f^2]} + 6r_n^\star + 14MB_n$$

with probability at least $1 - e^{-t}$. Applying Lemma 18 to this bound with the choice $\eta = 1$ immediately yields that

$$\begin{aligned}\mathbb{E}[f] &\le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2V_n \mathbb{E}_{\widehat{P}_n}[f^2] + 144M^2 V_n r_n^\star + 7V_n M \max\{M,1\}t/n} + 6r_n^\star + 14MB_n \\ &\le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2V_n \mathbb{E}_{\widehat{P}_n}[f^2]} + 12M\sqrt{V_n\left(r_n^\star + \frac{7\max\{M,1\}}{M}\frac{t}{n}\right)} + 6r_n^\star + 14MB_n\end{aligned}$$

for all $f \in \mathcal{F}$ with probability at least $1 - 2e^{-t}$. Subtracting and adding $(\mathbb{E}_{\widehat{P}_n}[f])^2$ to the second term, we have

$$\sqrt{2V_n \mathbb{E}_{\widehat{P}_n}[f^2]} = \sqrt{2V_n \mathrm{Var}_{\widehat{P}_n}(f) + 2V_n \mathbb{E}_{\widehat{P}_n}[f]^2} \le \sqrt{2V_n \mathrm{Var}_{\widehat{P}_n}(f)} + \sqrt{2V_n}\mathbb{E}_{\widehat{P}_n}[f],$$

where we have used that $f \ge 0$. We thus obtain

$$\begin{aligned}\mathbb{E}[f] &\le \left(1 + \sqrt{2V_n}\right)\mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2V_n \mathrm{Var}_{\widehat{P}_n}(f)} + 12M\sqrt{V_n\left(r_n^\star + \frac{7\max\{M,1\}}{M}\frac{t}{n}\right)} + 6r_n^\star + 14MB_n \\ &\le \left(1 + \sqrt{2V_n}\right)\mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2V_n \mathrm{Var}_{\widehat{P}_n}(f)} + 6MV_n + 6M\left(r_n^\star + \frac{7\max\{M,1\}t}{Mn}\right) + 6r_n^\star + 14MB_n,\end{aligned}$$

where the second inequality follows because $\sqrt{ab} \le \frac{1}{2}a + \frac{1}{2}b$ for $a, b \ge 0$. Recalling the bound (21), which implies $\rho \ge nV_n$, $\rho \ge n(r_n^\star + \frac{7\max\{M,1\}t}{Mn})$, and $\rho/n \ge 6r_n^\star + 14MB_n$, we obtain that

$$\mathbb{E}[f] \le \left(1 + \sqrt{\frac{2\rho}{n}}\right)\mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(f)} + \frac{13M\rho}{n}.$$

Theorem 1 implies $\mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(f)} \le \sup_{P: D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f(X)] + \frac{2M\rho}{n}$, so we immediately we arrive at

$$\mathbb{E}[f] \le \left(1 + 2\sqrt{\frac{2\rho}{n}}\right)\sup_{P: D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f(X)] + \left(13 + 4\sqrt{\frac{2\rho}{n}}\right)\frac{M\rho}{n}$$

for all $f \in \mathcal{F}$ with probability at least $1 - 2e^{-t}$. This is the first result (22).

To show the second result, we simply apply Bernstein's inequality and the concentration inequalities for the standard deviation in Lemma 11. For any fixed $f \in \mathcal{F}$, by Bernstein's inequality, we have

$$\mathbb{E}_{\widehat{P}_n}[f] \le \mathbb{E}[f] + \sqrt{\frac{2t\mathrm{Var}(f)}{n}} + \frac{2Mt}{3n}$$

with probability at least $1 - e^{-t}$. From Lemma 11, we have

$$\sqrt{\mathrm{Var}_{\widehat{P}_n}(f)} \le \sqrt{\mathrm{Var}(f)} + \sqrt{\frac{2tM^2}{n}}$$

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

We thus obtain that for any fixed f,

$$\sup_{P: D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(f)} \le \mathbb{E}[f] + \sqrt{\frac{2t}{n}\mathrm{Var}(f)} + \sqrt{\frac{2\rho}{n}\mathrm{Var}(f)} + \frac{2M\sqrt{\rho t}}{n} + \frac{2Mt}{3n}$$

with probability at least $1 - 2e^{-t}$. Noting that $\rho \ge 45Mt$ by assumption (21), so $\sqrt{\rho} + \sqrt{t} \le \sqrt{46\rho/45 + 45t} \le \sqrt{91\rho/45}$ and that always $2\sqrt{\rho t} \le 3\rho + \frac{1}{3}t$, we have that with probability at least $1 - 2e^{-t}$ that

$$\sup_{P: D_\phi(P\|\widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[f] \le \mathbb{E}[f] + \sqrt{\frac{91\rho}{45n}\mathrm{Var}(f)} + \frac{3M\rho}{n} + \frac{Mt}{n}.$$

Noting that we could take f to minimize the right hand side of the preceding expression and that $\widehat{f}$ minimizes $\sup_{P: D_\phi(P\|\widehat{P}_n) \le \rho/n} \mathbb{E}_P[f]$, we have the result (23).

We first show the claim for $g \in \mathcal{F}_{\mathrm{centered}} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$. To see the claim for $g \in \mathcal{F}_{\mathrm{centered}}$, let us fix $L \in \mathbb{N}$ to be chosen later, and for $l = 1, \ldots, L-1$ define the classes

$$\mathcal{F}_l := \left\{g \in \mathcal{F}_{\mathrm{centered}} : e^{-l} r < \mathbb{E}[g^2] \le e^{-(l-1)} r\right\}, \quad \mathcal{F}_L := \left\{g \in \mathcal{F}_{\mathrm{centered}} : \mathbb{E}[g^2] \le e^{-L} r\right\}$$

so that $\mathcal{F}_{\mathrm{centered}} = \cup_{l=1}^L \mathcal{F}_l$. Let $z > 0$ be such that $t \le z$. Applying Lemma 14 (with the choice $\alpha = \frac{1}{2}$) to $\mathcal{F}_l$ for each $l = 1, \ldots, L-1$, we have with probability at least $1 - e^{-t}$, for every $g \in \mathcal{F}_l$

$$\begin{aligned}\mathbb{E}[g] &\le \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2te^{-(l-1)}r}{n}} + 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_l)] + 5M\frac{t}{n} \\ &\le \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2et}{n}\mathbb{E}[g^2]} + 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_l)] + 5M\frac{t}{n}\end{aligned}$$

where in the last line we have used $e^{-l} r \le \mathbb{E}[g^2]$ for $g \in \mathcal{F}_l$. Similarly, applying Lemma 14 to $\mathcal{F}_L$, then with probability at least $1 - e^{-t}$, for every $g \in \mathcal{F}_L$

$$\begin{aligned}\mathbb{E}[g] &\le \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2te^{-L}r}{n}} + 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_L)] + 5M\frac{t}{n} \\ &\le \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2et}{n}\mathbb{E}[g^2]} + \sqrt{\frac{2te^{-L}r}{n}} + 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_L)] + 5M\frac{t}{n}.\end{aligned}$$

Taking a union bound, we have with probability at least $1 - Le^{-t}$, for every $g \in \mathcal{F}_{\mathrm{centered}}$

$$\mathbb{E}[g] \le \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2et}{n}\mathbb{E}[g^2]} + 3\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_{\mathrm{centered}})] + 5M\frac{t}{n} + \sqrt{\frac{2te^{-L}r}{n}}.$$

Noting that $\mathbb{E}[\mathfrak{R}_n(\mathcal{F}_{\mathrm{centered}})] \le 2\mathbb{E}[\mathfrak{R}_n(\mathcal{F})]$ by Jensen's inequality, we take $L = \left\lceil \log \frac{rn}{M^2 t} \right\rceil$ and map t to $t + \log L$ to obtain the lemma. The case when the roles of $\mathbb{E}[f]$ and $\mathbb{E}_{\widehat{P}_n}[f]$ are reversed follows similarly.

D.1. Proof of Lemma 17

Let $r \geq r_n^\star$ be an arbitrary but fixed value to be choosen later. Using this r, define the self-normalized class of functions

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

From the truncation by r, we have $\mathbb{E}[g^2] \leq r$ for all $g \in \mathcal{G}_r$. Lemma 16 implies that with probability at least $1 - e^{-t}$, uniformly over $g \in \mathcal{G}_r$

$$\mathbb{E}[g] \leq \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{\frac{2e}{n}\mathbb{E}[g^2]\left(t + \log\left\lceil \log \frac{n}{t} \right\rceil\right)} + 6\mathbb{E}[\mathfrak{R}_n(\mathcal{G}_r)] + \frac{7M}{n}\left(t + \log\left\lceil \log \frac{n}{t} \right\rceil\right). \quad (34)$$

Using the sub-root property of ψ_n and that $\psi_n(r_n^\star) = r_n^\star$, we have the inequality

$$\psi_n(r) = \sqrt{r}\psi_n(r)/\sqrt{r} \leq \sqrt{r}\psi_n(r_n^\star)/\sqrt{r_n^\star} = \sqrt{r r_n^\star}$$

for any $r \geq r_n^\star$, so

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

Using this upper bound in Eq. (34) and recalling the notation $B_n = \frac{1}{n}\left(t + \log\left\lceil \log \frac{n}{t} \right\rceil\right)$, we get

$$\mathbb{E}[g] \leq \mathbb{E}_{\widehat{P}_n}[g] + \sqrt{2eB_n\mathbb{E}[g^2]} + 6\sqrt{r_n^\star r} + 7MB_n. \quad (35)$$

Now, we return to choose the value r to optimize the bound (35). let r be the largest solution to $6\sqrt{r_n^\star r} + 7MB_n = 6r$. The following elementary lemma provides a bound on r.

Lemma 19 *Let x be the largest solution to $ax + b = \frac{x^2}{d}$ where $a, b, d > 0$. Then $a^2d^2 \leq x^2 \leq a^2d^2 + 2bd$.*

Proof From the quadratic formula, we have $x = \frac{1}{2}\left(ad + \sqrt{a^2d^2 + 4b}\right)$ from which the lower bound follows. From convexity of $z \mapsto z^2$ and $\sqrt{z_1 + z_2} \leq \sqrt{z_1} + \sqrt{z_2}$ for $z_1, z_2 > 0$, we obtain the upper bound. ■

Lemma 19 immediately yields

$$r_n^\star \leq r \leq r_n^\star + \frac{7MB_n}{3}.$$

For each $g \in \mathcal{G}_r$, there exists $f \in \mathcal{F}$ such that $g = \sqrt{\frac{r}{\mathbb{E}[f^2] \vee r}} f$. If $\mathbb{E}[f^2] \leq r$, we have $g = f$ and the bound (35) yields

$$\mathbb{E}[f] \leq \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2eB_n\mathbb{E}[f^2]} + 6r_n^\star + 14MB_n.$$

If $\mathbb{E}[f^2] > r$, rescaling g in the bound (35) and using the choice $6r = 6\sqrt{r_n^\star r} + 7MB_n$ yields

$$\begin{aligned}\mathbb{E}[f] &\leq \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2eB_n\mathbb{E}[f^2]} + 6\sqrt{r\mathbb{E}[f^2]} \\ &\leq \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{2eB_n\mathbb{E}[f^2]} + 6\sqrt{(r_n^\star + 7MB_n/3)\mathbb{E}[f^2]}\end{aligned}$$

instead. Combining the cases $\mathbb{E}[f^2] \lessgtr r$, we conclude that for all $f \in \mathcal{F}$,

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \left(\sqrt{2eB_n} + 6\sqrt{r_n^\star + 7MB_n/3}\right)\sqrt{\mathbb{E}[f^2]} + 6r_n^\star + 14MB_n$$

with probability at least $1 - e^{-t}$. Similarly, we can reverse the roles of $\mathbb{E}[f]$ and $\mathbb{E}_{\widehat{P}_n}[f]$ to get the second result.

D.2. Proof of Lemma 18

We frequently use the Rademacher contraction principle (Ledoux and Talagrand, 1991, Thm. 4.12) in what follows.

Lemma 20 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz. Then, for every class $\mathcal{G}$*

$$\mathbb{E}_\epsilon[\mathfrak{R}_n(\phi \circ \mathcal{G})] \le L\mathbb{E}_\epsilon[\mathfrak{R}_n(\mathcal{G})]$$

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

As in Section D.1, define the self-normalized functions in $\mathcal{F}$

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

where $r \ge r_n^\star$ will be choosen later. Let $\mathcal{G}_r^2 = \{g^2 : g \in \mathcal{G}_r\}$. From the truncation by r, we have that for all $g^2 \in \mathcal{G}_r^2$, $\mathrm{Var}(g^2) \le \mathbb{E}[g^4] \le M^2\mathbb{E}[g^2] \le M^2 r$. Let $c_1 = 3$ and $c_2 = \frac{7}{3}$. Then by Lemma 14 applied to $\mathcal{G}_r^2$, with probability at least $1 - e^{-t}$, for every $g \in \mathcal{G}_r$

$$\begin{aligned}
\mathbb{E}[g^2] &\le \mathbb{E}_{\widehat{P}_n}[g^2] + c_1\mathbb{E}[\mathfrak{R}_n(\mathcal{G}_r^2)] + M\sqrt{\frac{2rt}{n}} + c_2\frac{M^2 t}{n} \\
&\overset{(a)}{\le} \mathbb{E}_{\widehat{P}_n}[g^2] + 2c_1 M\mathbb{E}[\mathfrak{R}_n(\mathcal{G}_r)] + M\sqrt{\frac{2rt}{n}} + \frac{c_2 M^2 t}{n} \\
&\overset{(b)}{\le} \mathbb{E}_{\widehat{P}_n}[g^2] + 2c_1 M\sqrt{r r_n^\star} + M\sqrt{\frac{2rt}{n}} + \frac{c_2 M^2 t}{n} \qquad (36)
\end{aligned}$$

where in step (a) we used the contraction principle (Lemma 20) and that $x \mapsto x^2$ is $2M$-Lipschitz on $[-M, M]$, and in step (b), we used that $\psi_n(r) \le \sqrt{r r_n^\star}$ as in the proof of Lemma 17 in Section D.1.

Let $A = 2c_1 M\sqrt{r_n^\star} + M\sqrt{\frac{2t}{n}}$ and $D = \frac{c_2 M^2 t}{n}$. For any fixed $K > 1$, choose r to be the largest solution to $A\sqrt{r} + D = \frac{r}{K}$ so that the bound (36) becomes

$$\mathbb{E}[g^2] \le \mathbb{E}_{\widehat{P}_n}[g^2] + \frac{r}{D}.$$

From Lemma 19, we have

$$K^2A^2 \le r \le K^2A^2 + 2KD$$

and in particular, $r \ge K^2A^2 \ge r_n^\star$. For each $g \in \mathcal{G}_r$, there exists $f \in \mathcal{F}$ such that $g = \sqrt{\frac{r}{\mathbb{E}[f^2] \vee r}} f$. If $\mathbb{E}[f^2] \le r$, rescaling the inequality (36) and using the upper bound on r, we obtain

$$\mathbb{E}[f^2] \le \mathbb{E}_{\widehat{P}_n}[f^2] + \frac{r}{K} \le \mathbb{E}_{\widehat{P}_n}[f^2] + KA^2 + 2D.$$

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

$$\mathbb{E}[f^2] \le \mathbb{E}_{\widehat{P}_n}[f^2] + \frac{\mathbb{E}[f^2]}{K}.$$

Combining the two cases, we obtain

$$\mathbb{E}[f^2] \le \frac{K}{K-1}\mathbb{E}_{\widehat{P}_n}[f^2] + KA^2 + 2D.$$

Noting that $A \le 2\left(4c_1^2M^2r_n^\star + 2\frac{M^2t}{n}\right)$ by convexity, we have the first result once we replace K with $\eta = K - 1 > 0$. The second result similarly follows by reversing the roles of $\mathbb{E}[f]$ and $\mathbb{E}_{\widehat{P}_n}[f]$ in the above argument.

Appendix E. Proof of Theorem 8

Recall our shorthand notation that $\pi(\theta) = \operatorname{argmin}_{\theta^* \in S_\star}\{\|\theta - \theta^*\|_2\}$ denotes the Euclidean projection of θ onto $S_\star$, which is a closed convex set. Define also the localized empirical deviation function

$$\Delta_n(\theta) := \mathbb{E}\left[\ell(\theta; X) - \ell(\pi(\theta); X)\right] - \mathbb{E}_{\widehat{P}_n}\left[\ell(\theta; X) - \ell(\pi(\theta); X)\right]. \tag{37}$$

We begin with the following

Claim 21 *If $\widehat{S}_\star^\epsilon \not\subset S_\star^{2\epsilon}$, then*

$$\sup_{\theta \in S_\star^{2\epsilon}} \left\{\Delta_n(\theta) + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X))}\right\} \ge \epsilon. \tag{38}$$

Deferring the proof of the claim, let us prove the theorem. First, the growth condition (26) shows that

$$S_\star^{2\epsilon} \subset \left\{\theta \in \Theta : \|\theta - \pi(\theta)\|_2 \le \left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}}\right\} = \left\{\theta \in \Theta : \operatorname{dist}(\theta, S_\star) \le \left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}}\right\}.$$

Therefore, we have for all $\theta \in S_\star^{2\epsilon}$ that

$$\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X)) \le L^2 \operatorname{dist}(\theta, S_\star)^2 \le L^2\left(\frac{2\epsilon}{\lambda}\right)^{\frac{2}{\gamma}},$$

and so by the assumption (27) that $\epsilon \ge (\frac{8L^2\rho}{n})^{\frac{\gamma}{2(\gamma-1)}}(\frac{2}{\lambda})^{\frac{1}{\gamma-1}}$, we have

$$\sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X))} \le L\sqrt{\frac{2\rho}{n}}\left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}} \le \frac{\epsilon}{2}.$$

In particular, if the event (38) holds then

$$\sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta) \ge \frac{\epsilon}{2},$$

and recalling the definition (37) of Δ_n, it then follows that

$$\mathbb{P}\left(\widehat{S}_\star^\epsilon \not\subset S_\star^{2\epsilon}\right) \le \mathbb{P}\left(\sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta) \ge \frac{\epsilon}{2}\right). \tag{39}$$

To bound the probability (39), we use standard bounded difference and symmetrization arguments (e.g. Boucheron et al., 2013, Theorem 6.5). Letting $f(X_1, \ldots, X_n) := \sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta)$, the function f satisfies bounded differences:

$$\begin{aligned}
&\sup_{x, x' \in \mathcal{X}} |f(X_1, \cdots, X_{j-1}, x, X_{j+1}, \cdots, X_n) - f(X_1, \cdots, X_{j-1}, x', X_{j+1}, \cdots, X_n)| \\
&\le \sup_{x, x' \in \mathcal{X}} \sup_{\theta \in S_\star^{2\epsilon}} \left| \frac{1}{n}(\ell(\theta; x) - \ell(\pi(\theta); x)) - \frac{1}{n}(\ell(\theta; x') - \ell(\pi(\theta); x')) \right| \\
&\le \frac{2L}{n} \sup_{\theta \in S_\star^{2\epsilon}} \operatorname{dist}(\theta, S_\star) \le \frac{2L}{n} \left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}}
\end{aligned}$$

for $j = 1, \ldots, n$. Using the standard symmetrization inequality $\mathbb{E}[\sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta)] \le 2\mathbb{E}[\mathfrak{R}_n(S_\star^{2\epsilon})]$ and the bounded differences inequality (Boucheron et al., 2013, Theorem 6.5), we have

$$\mathbb{P}\left(\sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta) \ge 2\mathbb{E}[\mathfrak{R}_n(S_\star^{2\epsilon})] + t\right) \le \exp\left(-\frac{nt^2}{2L^2}\left(\frac{\lambda}{2\epsilon}\right)^{\frac{2}{\gamma}}\right)$$

for all $t \ge 0$. Letting $u = \frac{nt^2}{2L^2}\left(\frac{\lambda}{2\epsilon}\right)^{\frac{2}{\gamma}}$ above and recalling the assumption (27) upper bounding $\mathbb{E}[\mathfrak{R}_n(S_\star^{2\epsilon})]$, we have $\mathbb{P}(\sup_{\theta \in S_\star^{2\epsilon}} \Delta_n(\theta) \ge \frac{\epsilon}{2}) \le e^{-u}$. The theorem follows from the bound (39).

Proof of Claim 21 If $\widehat{S}_\star^\epsilon \not\subset S_\star^{2\epsilon}$, then certainly it is the case that there is some $\theta \in \Theta \setminus S_\star^{2\epsilon}$ such that

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

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

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

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 $S_\star$ at $\pi(\theta)$; cf. (Hiriart-Urruty and Lemaréchal, 1993, Prop. III.5.3.3)) that $\pi(t\theta + (1-t)\pi(\theta)) = \pi(\theta)$. Thus, choosing t appropriately, there exists $\theta' \in \operatorname{bd} S_\star^{2\epsilon}$ with $\theta' = t\theta + (1-t)\pi(\theta)$, $\pi(\theta') = \pi(\theta)$, and $R_n(\theta', \mathcal{P}_n) \le R_n(\pi(\theta'), \mathcal{P}_n) + \epsilon$.

Adding and subtracting the risk $R(\theta)$ and $R(\pi(\theta))$, we have that for some $\theta \in \operatorname{bd} S_\star^{2\epsilon}$ that

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

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

$$\begin{aligned}
\epsilon &\le \sup_{\theta \in S_\star^{2\epsilon}} \left\{R(\theta) - R_n(\theta, \mathcal{P}_n) - (R(\pi(\theta)) - R_n(\pi(\theta), \mathcal{P}_n))\right\} \\
&\le \sup_{\theta \in S_\star^{2\epsilon}} \sup_{P : D_\phi(P \| \widehat{P}_n) \le \rho/n} \left\{R(\theta) - R(\pi) + \mathbb{E}_P\left[\ell(\pi(\theta); X) - \ell(\theta; X)\right]\right\}.
\end{aligned}$$

Applying the upper bound in inequality (10) of Theorem 1 gives the claim.

Appendix F. Proof of Theorem 10

We begin by establishing a few technical lemmas, after which the proof of the theorem follows essentially standard arguments in asymptotics. To prove Theorem 10, we first show that (eventually) we have the exact expansion

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

for all θ in a neighborhood of $\theta^\star$. As in the proof of Theorem 1, this exact equality holds once there is suitable variability in the values $\ell(\theta, X_i)$ over $i = 1, \ldots, n$, however, we require a bit more care as the values $\ell(\theta, X_i)$ may be unbounded below and above. Heuristically, however, assuming that we have this exact expansion and that $\widehat{\theta}_n^{\rm rob} - \theta^\star = O_P(n^{-\frac{1}{2}})$, then we can write the expansions

$$\begin{aligned}
0 &= \nabla_\theta R_n(\widehat{\theta}_n^{\rm rob}, \mathcal{P}_n) \\
&= \nabla \frac{1}{n} \sum_{i=1}^n \ell(\theta^\star, X_i) + \nabla^2 \left(\frac{1}{n} \sum_{i=1}^n \ell(\theta^\star, X_i) \right) (\widehat{\theta}_n^{\rm rob} - \theta^\star) + \nabla \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{\theta}_n^{\rm rob}, X))}{n}} + o_P(n^{-\frac{1}{2}}) \\
&= \frac{1}{n} \sum_{i=1}^n \nabla \ell(\theta^\star, X_i) + \nabla^2 R(\theta^\star)(\widehat{\theta}_n^{\rm rob} - \theta^\star) + \nabla \sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta^\star, X))}{n}} + o_P(n^{-\frac{1}{2}}).
\end{aligned}$$

Multiplying by $\sqrt{n}$ and solving for $\widehat{\theta}_n^{\rm rob}$ in the preceding expression, computing $\nabla \sqrt{\mathrm{Var}(\ell(\theta^\star, X))}$ then yields the theorem.

The remainder of the proof makes this heuristic rigorous, and the outline is as follows:

1. We show that there is a uniform expansion of the form (12) in a neighborhood of $\theta^\star$. (See Section F.1.)

2. Using the uniform expansion, we can then leverage standard techniques for asymptotic analysis of finite-dimensional estimators (see, e.g. van der Vaart and Wellner (1996) or Lehmann and Casella (1998)), which proceed by performing a Taylor expansion of the objective in a neighborhood of the optimum and using local asymptotic normality arguments. (See Section F.2.)

F.1. The uniform variance expansion

To lighten notation, we define a few quantities similar to those used in the proof of Theorem 1. Let

$$Z(\theta) := \ell(\theta, X) - \mathbb{E}[\ell(\theta, X)]$$

be the deviation of $\ell(\theta, X)$ around its mean (the risk), and similarly let $Z_i(\theta)$ be the version of this quantity for observation X_i. In addition, let $s_n^2(\theta) = \mathrm{Var}_{\widehat{P}_n}(Z(\theta))$ be the empirical variance of $Z(\theta)$, which is identical to the empirical variance of $\ell(\theta, X)$.

Now, recall the problem

$$\mathop{\mathrm{maximize}}_P \; \mathbb{E}_P[Z(\theta)] \quad \text{subject to } D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n},$$

and for each $\theta \in \Theta$, let $p(\theta) = \operatorname{argmax}_{p \in \mathcal{P}_n} \sum_{i=1}^n p_i Z_i(\theta)$ be the solution (probability) vectors. Following expression (9) we see for any $\epsilon \geq 0$ that

$$\min_{i \in [n]} \frac{\sqrt{2\rho}(Z_i(\theta) - \overline{Z}(\theta))}{\sqrt{n} s_n(\theta)} \geq -1 \quad \text{for all } \theta \in \theta^\star + \epsilon \mathbb{B}$$

is sufficient for the exact variance expansion to hold. We now show that this is indeed likely. Let $\epsilon > 0$ be small enough that Assumption A holds, that is, the random Lipschitz function $L(X)$ satisfies $|\ell(\theta, x) - \ell(\theta', x)| \leq L(x)\|\theta - \theta'\|$ for $\theta, \theta' \in \theta^\star + \epsilon \mathbb{B}$. Then because

$$\begin{aligned} \left|\sqrt{n} s_n(\theta) - \sqrt{n} s_n(\theta')\right| & \leq \sup_{u : \|u\|_2 \leq 1} \sum_{i=1}^n u_i \left(\ell(\theta, X_i) - \ell(\theta', X_i)\right) \\ & \leq \sup_{u : \|u\|_2 \leq 1} \sum_{i=1}^n u_i L(X_i) \left\|\theta - \theta'\right\| \leq \sqrt{\sum_{i=1}^n L^2(X_i)} \left\|\theta - \theta'\right\| \end{aligned}$$

so $\theta \mapsto s_n(\theta)$ is $\sqrt{\frac{1}{n}\sum_{i=1}^n L(X_i)^2}$-Lipschitz for $\theta \in \theta^\star + \epsilon \mathbb{B}$, we have

$$\inf_{\theta \in \theta^\star + \epsilon \mathbb{B}} \min_{i \in [n]} \left\{ \frac{\sqrt{2\rho}(Z_i(\theta) - \overline{Z}(\theta))}{\sqrt{n} s_n(\theta)} \right\} \geq \min_{i \in [n]} \frac{\sqrt{2\rho}(Z_i(\theta^\star) - \overline{Z}(\theta^\star) - 2\epsilon L(X_i))}{\sqrt{n\left(s_n(\theta^\star) - \epsilon\sqrt{\frac{1}{n}\sum_{j=1}^n L(X_j)^2}\right)}}.$$

Summarizing our development thus far, we have the following lemma.

Lemma 22 *Let the conditions of the previous paragraph hold. Then*

$$\min_{i \in [n]} \left\{ \sqrt{2\rho}(Z_i(\theta^\star) - \overline{Z}(\theta^\star) - 2\epsilon L(X_i)) \right\} \geq \sqrt{n} \sqrt{s_n(\theta^\star) - \epsilon \left(\frac{1}{n}\sum_{i=1}^n L(X_i)^2\right)^{\frac{1}{2}}}$$

implies that

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2\rho}{n} \operatorname{Var}_{\widehat{P}_n}(\ell(\theta, X))} \quad \textit{for all } \theta \in \theta^\star + \epsilon \mathbb{B}.$$

Now, we use the following standard result to show that the conditions of Lemma 22 eventually hold with probability one.

Lemma 23 (Owen (Owen, 1990), Lemma 3) *Let Y_i be independent random variables with $\sup_i \mathbb{E}[Y_i^2] < \infty$. Then $n^{-\frac{1}{2}} \max_{1 \leq i \leq n} |Y_i| \stackrel{a.s.}{\to} 0$.*

Based on Lemma 23 and the strong law of large numbers, we see immediately that

$$\frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} |Z_i(\theta^\star)| \stackrel{a.s.}{\to} 0, \quad \text{and} \quad \frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} L(X_i) \stackrel{a.s.}{\to} 0,$$

because $\mathbb{E}[Z(\theta^\star)^2] < \infty$ and $\mathbb{E}[L(X_i)^2] < \infty$. Applying the strong law of large numbers to obtain

$$s_n(\theta^\star) \stackrel{a.s.}{\to} \sqrt{\operatorname{Var}(\ell(\theta^\star, X))} \quad \text{and} \quad \sqrt{\frac{1}{n}\sum_{i=1}^n L(X_i)^2} \stackrel{a.s.}{\to} \sqrt{\mathbb{E}[L(X)^2]},$$

we see immediately that for small enough $\epsilon > 0$, the condition of Lemma 22 holds eventually with probability 1. That is, the following uniform expansion holds.

Lemma 24 *There exists $\epsilon > 0$ such that, with probability 1, there exists an N (which may be random) such that $n \geq N$ implies*

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta, X))}{n}} \quad \textit{for all } \theta \in \theta^\star + \epsilon\mathbb{B}.$$

F.2. Asymptotics and Taylor expansions

Let $\mathcal{E}_{n,\mathrm{exact}}$ be the event that the exact variance expansion of Lemma 24 occurs for $\theta \in \theta^\star + \epsilon\mathbb{B}$. Now that we know that $\mathbb{P}(\mathcal{E}_{n,\mathrm{exact}} \text{ eventually}) = 1$, we may perform a few asymptotic expansions of the variance-regularized objective to provide the convergence guarantees specified by the theorem. We use the following lemma.

Lemma 25 *Let the conditions of the theorem hold. If*

$$\widehat{\theta}_n^{\mathrm{rob}} \in \operatorname*{argmin}_\theta R_n(\theta, \mathcal{P}_n) \quad \textit{then} \quad \widehat{\theta}_n^{\mathrm{rob}} \stackrel{a.s.}{\to} \theta^\star. \tag{40}$$

The proof is standard, but for completeness we include it in Section G.3.

By combining Lemmas 24 and 25, we see that with probability 1, for any $\epsilon > 0$, we eventually have both

$$\|\widehat{\theta}_n^{\mathrm{rob}} - \theta^\star\|_2 < \epsilon \quad \text{and} \quad R_n(\widehat{\theta}_n^{\mathrm{rob}}, \mathcal{P}_n) = \mathbb{E}_{\widehat{P}_n}[\ell(\widehat{\theta}_n^{\mathrm{rob}}, X)] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X))}.$$

Assume for the remainder of the argument that both of these conditions hold. Standard results on subdifferentiability of maxima of collections of convex functions (Hiriart-Urruty and Lemaréchal, 1993, Chapter X) give that $R_n(\theta, \mathcal{P}_n)$ is differentiable near $\theta^\star$, and thus

$$\begin{aligned} 0 &= \nabla R_n(\widehat{\theta}_n^{\mathrm{rob}}, \mathcal{P}_n) = \mathbb{E}_{\widehat{P}_n}[\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X)] + \nabla \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X))} \\ &= \frac{1}{n}\sum_{i=1}^n \nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X_i) + \sqrt{\frac{2\rho}{n}} \frac{\mathbb{E}_{\widehat{P}_n}\left[(\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X) - \mathbb{E}_{\widehat{P}_n}[\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X)])(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X) - \mathbb{E}_{\widehat{P}_n}[\ell(\widehat{\theta}_n^{\mathrm{rob}}, X)])\right]}{\sqrt{\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X))}}. \end{aligned} \tag{41}$$

Because $\widehat{\theta}_n^{\mathrm{rob}} \stackrel{a.s.}{\to} \theta^\star$, by the continuous mapping theorem and local uniform convergence of the empirical expectations $\mathbb{E}_{\widehat{P}_n}[\cdot]$ to $\mathbb{E}[\cdot]$, the second term of expression (41) satisfies

$$\frac{\mathbb{E}_{\widehat{P}_n}\left[(\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X) - \mathbb{E}_{\widehat{P}_n}[\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X)])(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X) - \mathbb{E}_{\widehat{P}_n}[\ell(\widehat{\theta}_n^{\mathrm{rob}}, X)])\right]}{\sqrt{\mathrm{Var}_{\widehat{P}_n}(\ell(\widehat{\theta}_n^{\mathrm{rob}}, X))}} = \underbrace{\frac{\mathrm{Cov}(\nabla \ell(\theta^\star, X), \ell(\theta^\star, X))}{\sqrt{\mathrm{Var}(\ell(\theta^\star, X))}}}_{=: b(\theta^\star)} + o_P(1).$$

For simplicity, we let $b(\theta^\star)$ denote the final term, which we shall see becomes an asymptotic bias. Thus, performing a Taylor expansion of the terms $\nabla \ell(\widehat{\theta}_n^{\mathrm{rob}}, X_i)$ around $\theta^\star$ in equality (41), there exist (random) error matrices $E_n(X_i)$, where $\|E_n(X_i)\| \leq H(X_i)\|\widehat{\theta}_n^{\mathrm{rob}} - \theta^\star\|$

by Assumption A, such that

$$0 = \mathbb{E}_{\widehat{P}_n}[\nabla \ell(\theta^\star, X)] + \frac{1}{n}\sum_{i=1}^n \left(\nabla^2 \ell(\theta^\star, X_i) + E_n(X_i)\right)(\widehat{\theta}_n^{\rm rob} - \theta^\star) + \sqrt{\frac{2\rho}{n}}(b(\theta^\star) + o_P(1))$$
$$= \mathbb{E}_{\widehat{P}_n}[\nabla \ell(\theta^\star, X)] + \left(\nabla^2 R(\theta^\star) + o_P(1)\right)(\widehat{\theta}_n^{\rm rob} - \theta^\star) + \sqrt{\frac{2\rho}{n}}(b(\theta^\star) + o_P(1)).$$

Multiplying both sides by $\sqrt{n}$, using that $\nabla^2 R(\theta^\star) + o_P(1)$ is eventually invertible, and applying the continuous mapping theorem, we have

$$\sqrt{n}(\widehat{\theta}_n^{\rm rob} - \theta^\star) = -(\nabla^2 R(\theta^\star) + o_P(1))^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla \ell(\theta^\star, X_i) - \sqrt{2\rho} b(\theta^\star) + o_P(1).$$

The first term on the right side of the above display converges in distribution to a $\mathsf{N}(0, \Sigma)$ distribution, where

$$\Sigma = (\nabla^2 R(\theta^\star))^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star, X))(\nabla^2 R(\theta^\star))^{-1},$$

so that

$$\sqrt{n}(\widehat{\theta}_n^{\rm rob} - \theta^\star) \stackrel{d}{\to} \mathsf{N}\left(-\sqrt{2\rho}\, b(\theta^\star), \Sigma\right)$$

as claimed in the theorem statement.

Appendix G. Proofs of Technical Lemmas

G.1. Proof of Inequality (24)

Define the Gaussian complexity

$$\mathfrak{G}_n(\{\ell \circ \mathcal{H}\}_{\le r}) := \mathbb{E}\left[\sup_{h \in \mathbb{B}_{\mathcal{H}}, c \in [0,1]} \sum g_i c \ell(h(x_i), y_i) \mid \mathbb{E}[\ell(h(X), Y)^2] \le r/c^2\right], \tag{42}$$

where $g_i \stackrel{\rm iid}{\sim} \mathsf{N}(0,1)$ (here we recall the standard result (Bartlett and Mendelson, 2002) that Gaussian complexity upper bounds Rademacher complexities up to a constant). Now, the set $h - h^\star$ such that $h \in \mathbb{B}_{\mathcal{H}}$ is contained in $2\mathbb{B}_{\mathcal{H}}$, which is convex. Moreover, we have $\mathbb{E}[\ell(h(X), Y)^2] = \mathbb{E}[(h(X) - h^\star(X))^2] + \sigma^2$, and so we have for any c that

$$\{h \in \mathbb{B}_{\mathcal{H}} \mid c^2 \mathbb{E}[\ell(h(X), Y)^2] \le r\} \subset \{h \in \mathbb{B}_{\mathcal{H}} \mid \mathbb{E}[(h(X) - h^\star(X))^2] \le r/c^2\},$$

and $\mathbb{E}[\ell(h(X), Y)^2] \le r/c^2$ also implies $\sigma^2 \le r/c^2$. Returning to expression (42) and enlarging the sets over which we take suprema, we thus obtain

$$\mathfrak{G}_n(\ell \circ \mathcal{H}) \le \mathbb{E}\left[\sup_{h \in \mathbb{B}_{\mathcal{H}}, c_1, c_2 \in [0,1]} \sum_{i=1}^n g_i |c_1(h(x_i) - h^\star(x_i)) - c_2 \xi_i| \mid \mathbb{E}[(h(X) - h^\star(X))^2] \le \frac{r}{c_1^2}, \sigma^2 \le \frac{r}{c_2^2}\right]$$
$$\le \mathbb{E}\left[\sup_{f \in 2\mathbb{B}_{\mathcal{H}}, c \in [0,1]} \sum_{i=1}^n g_i |f(x_i) - c\xi_i| \mid \mathbb{E}[f(X)^2] \le r, \sigma^2 \le r/c^2\right],$$

where we have used that $h - h^\star \in 2\mathbb{B}_{\mathcal{H}}$ and that the set $\mathbb{B}_{\mathcal{H}}$ is convex to obtain the second inequality. We now upper bound the final display using the classical Sudakov-Fernique comparison theorem (e.g. Chatterjee, 2005). Indeed, define the two Gaussian processes indexed by $f \in \mathcal{H}$ and $c \in [0,1]$ by $Y_{f,c} = \sum_{i=1}^n g_i |f(x_i) - c\xi_i|$ and $Z_{f,c} = \sum_{i=1}^n g_i f(x_i) + c \sum_{i=1}^n w_i \xi_i$, where $g_i \stackrel{\rm iid}{\sim} \mathsf{N}(0,1)$ and $w_i \stackrel{\rm iid}{\sim} \mathsf{N}(0,1)$. Then we have for any $f_1, f_2 \in \mathcal{H}$ and $c_1, c_2 \in [0,1]$ that

$$\begin{aligned}
\mathbb{E}[(Y_{f_1,c_1} - Y_{f_2,c_2})^2] &= \sum_{i=1}^n \left(|f_1(x_i) - c_1\xi_i| - |f_2(x_i) - c_2\xi_i|\right)^2 \\
&\le \sum_{i=1}^n \left(f_1(x_i) - f_2(x_i) + (c_2 - c_1)\xi_i\right)^2 \\
&\le 2\sum_{i=1}^n (f_1(x_i) - f_2(x_i))^2 + 2(c_2 - c_1)^2 \sum_{i=1}^n \xi_i^2.
\end{aligned}$$

Moreover, $\mathbb{E}[(Z_{f_1,c_1} - Z_{f_2,c_2})^2] = \sum_{i=1}^n (f_1(x_i) - f_2(x_i))^2 + (c_1 - c_2)^2 \sum_{i=1}^n \xi_i^2$. Thus, the Sudakov-Fernique inequality guarantees that $\mathbb{E}[\sup_{f,c} Y_{f,c}] \le \sqrt{2}\mathbb{E}[\sup_{f,c} Z_{f,c}]$, and

$$\mathfrak{G}_n(\ell \circ \mathcal{H}) \lesssim \mathbb{E}\left[\sup_{f \in 2\mathbb{B}_{\mathcal{H}}} \sum_{i=1}^n g_i f(x_i) \mid \mathbb{E}[f(X)^2] \le r\right] + \mathbb{E}\left[\sup_{c \in [0,1]} c \sum_{i=1}^n w_i \xi_i \mid c^2\sigma^2 \le r\right].$$

The last term in the expression has bound $\sqrt{nr}$ by Jensen's inequality and the relaxation that $c \in [-1,1]$. For the first term, Mendelson (2003, Thm. 2.1) shows that for RKHS with kernel eigenvalues $\lambda_1, \lambda_2, \ldots$, we have

$$\mathbb{E}\left[\sup_{f \in 2\mathbb{B}_{\mathcal{H}}} \sum_{i=1}^n g_i f(X_i) \mid \mathbb{E}[f(X)^2] \le r\right] \lesssim \sqrt{n} \left(\sum_{j=1}^\infty \min\{\lambda_j, r\}\right)^{\frac{1}{2}},$$

which yields our desired claim (24).

G.2. Proof of Lemma 7

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

$$\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)] = \frac{1}{n}\left[N_{-1}|\theta + 1| + N_1|\theta - 1| + N_0|\theta| - (n - N_0)\right],$$

because $N_1 + N_{-1} + N_0 = n$. In particular, we find that the empirical risk minimizer θ satisfies

$$\widehat{\theta}_n^{\rm erm} := \operatorname*{argmin}_{\theta \in \mathbb{R}} \mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)] = \begin{cases} 1 & \text{if } N_1 > N_0 + N_{-1} \\ -1 & \text{if } N_{-1} > N_0 + N_1 \\ \in [-1,1] & \text{otherwise.} \end{cases}$$

On the events $N_1 > N_{-1} + N_0$ or $N_{-1} > N_0 + N_1$, which are disjoint, then, we have

$$R(\widehat{\theta}_n^{\rm erm}) = \delta = R(\theta^\star) + \delta.$$

Let us give a lower bound on the probability of this event. Noting that marginally $N_1 \sim \mathsf{Bin}(n, \frac{1-\delta}{2})$ 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

$$\mathbb{P}\left(N_1 > \frac{n}{2}\right) = \mathbb{P}\left(\mathsf{Bin}\left(n, \frac{1-\delta}{2}\right) > \frac{n}{2}\right) = \mathbb{P}\left(\mathsf{Bin}\left(n, \frac{1+\delta}{2}\right) < \frac{n}{2}\right).$$

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 (2013) show that

$$\mathbb{P}\left(N_1 \geq \frac{n}{2}\right) \geq \Phi\left(-\sqrt{2n D_{\rm kl}\left(\frac{1}{2} \| \frac{1+\delta}{2}\right)}\right)$$

where $D_{\rm kl}\left(p \| q\right) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ denotes the binary KL-divergence. We have by standard bounds on the KL-divergence (Tsybakov, 2009, Lemma 2.7) that $D_{\rm kl}(\frac{1}{2} \| \frac{1+\delta}{2}) \leq \frac{\delta^2}{2(1-\delta^2)}$, 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

$$\mathbb{P}\left(N_1 = \frac{n}{2}\right) = 2^{-n} \binom{n}{n/2} (1-\delta^2)^{n/2} \leq (1-\delta^2)^{n/2} \sqrt{\frac{2}{\pi n}},$$

where the inequality uses that $\binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi n}}$ by Stirling's approximation. Summarizing, we find that

$$\mathbb{P}\left(N_1 > \frac{n}{2} \text{ or } N_{-1} > \frac{n}{2}\right) \geq 2\Phi\left(-\sqrt{\frac{n\delta^2}{1-\delta^2}}\right) - (1-\delta^2)^{n/2} \sqrt{\frac{8}{\pi n}}.$$

G.3. Proof of Lemma 25

Under the conditions of the theorem, the compactness of $\theta^\star + \epsilon\mathbb{B}$ guarantees that

$$\sup_{\theta \in \theta^\star + \epsilon\mathbb{B}} |\mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)] - R(\theta)| \stackrel{a.s.}{\to} 0,$$

as the functions $\theta \mapsto \ell(\theta, x)$ are Lipschitz in a neighborhood of $\theta^\star$ by Assumption A. Similarly,

$$\sup_{\theta \in \theta^\star + \epsilon\mathbb{B}} \left|\mathrm{Var}_{\widehat{P}_n}(\ell(\theta, X)) - \mathrm{Var}(\ell(\theta, X))\right| \stackrel{a.s.}{\to} 0,$$

using the local Lipschitzness of $\nabla^2 \ell$. (See, for example, the Glivenko-Cantelli results in Chapters 2.4–2.5 of van der Vaart and Wellner (1996).) Thus, using the two-sided bounds (10) of Theorem 1, we have that

$$\begin{aligned}
&\sup_{\theta \in \theta^\star + \epsilon\mathbb{B}} |R_n(\theta, \mathcal{P}_n) - R(\theta)| \\
&\leq \sup_{\theta \in \theta^\star + \epsilon\mathbb{B}} \left|\mathbb{E}_{\widehat{P}_n}[\ell(\theta, \mathcal{P}_n)] - R(\theta)\right| + \sqrt{\frac{2\rho}{n}} \sup_{\theta \in \theta^\star + \epsilon\mathbb{B}} \sqrt{\mathrm{Var}_{\widehat{P}_n}(\ell(\theta, X))} \stackrel{a.s.}{\to} 0.
\end{aligned}$$

Now, we use the fact that $\nabla^2 R(\theta^\star) \succ 0$, and that $\theta \mapsto \nabla^2 R(\theta)$ is continuous in a neighborhood of $\theta^\star$. Fix $\epsilon > 0$ small enough that the preceding uniform convergence guarantees hold over $\theta^\star + 2\epsilon\mathbb{B}$ and $\nabla^2 R(\theta) \succeq \lambda I$ for some $\lambda > 0$ and all $\theta \in \theta^\star + 2\epsilon\mathbb{B}$. Let $\theta \notin \theta^\star + \epsilon\mathbb{B}$, but $\theta \in \theta^\star + 2\epsilon\mathbb{B}$. Then for sufficently large n, we have that

$$\begin{aligned} R_n(\theta, \mathcal{P}_n) \geq \mathbb{E}_{\widehat{P}_n}[\ell(\theta, X)] &\stackrel{(i)}{\geq} R(\theta) - \frac{\lambda}{4}\epsilon^2 \\ &\stackrel{(ii)}{\geq} R(\theta^\star) + \frac{\lambda}{2}\left\|\theta - \theta^\star\right\|_2^2 - \frac{\lambda}{4}\epsilon^2 \stackrel{(iii)}{\geq} R(\theta^\star) + \frac{\lambda}{4}\epsilon^2 \\ &\stackrel{(iv)}{\geq} \mathbb{E}_{\widehat{P}_n}[\ell(\theta^\star, X)] + \frac{\lambda}{4}\epsilon^2 - \frac{\lambda}{8}\epsilon^2 = \mathbb{E}_{\widehat{P}_n}[\ell(\theta^\star, X)] + \frac{\lambda}{8}\epsilon^2, \end{aligned}$$

where inequalities (i) and (iv) follow from the uniform convergence guarantee, inequality (ii) from the strong convexity of R near $\theta^\star$, and (iii) because $\left\|\theta - \theta^\star\right\|_2 \geq \epsilon$. Finally, we have that

$$\mathbb{E}_{\widehat{P}_n}[\ell(\theta^\star, X)] \geq R_n(\theta^\star, \mathcal{P}_n) - \underbrace{\sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(\ell(\theta^\star, X))}}_{\stackrel{a.s.}{\to} 0},$$

so that eventually $R_n(\theta, \mathcal{P}_n) > R_n(\theta^\star, \mathcal{P}_n)$ for all $\theta \in \theta^\star + 2\epsilon\mathbb{B} \setminus \epsilon\mathbb{B}$. By convexity, then this inequality holds for all $\theta \notin \theta^\star + \epsilon\mathbb{B}$. Thus if $\widehat{\theta}_n^{\rm rob} \in \operatorname{argmin}_\theta R_n(\theta, \mathcal{P}_n)$, then for any $\epsilon > 0$ we must eventually have $\|\widehat{\theta}_n^{\rm rob} - \theta^\star\|_2 < \epsilon$.

Appendix H. Efficient solutions to computing the robust expectation

In this appendix, we give a detailed description of the procedure we use to compute the supremum problem (8). In particular, our procedure requires time $O(n \log n + \log \frac{1}{\epsilon} \log n)$, where ϵ 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

$$\text{minimize } \; p^\top z \quad \text{subject to } \; \frac{1}{2n}\left\|np - \mathbf{1}\right\|_2^2 \leq \rho, \; p \geq 0, \; p^\top \mathbf{1} = 1.$$

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 $\frac{1}{2}\|p - \frac{1}{n}\mathbf{1}\|_2^2 \leq \frac{\rho}{n}$ and performing the standard min-max swap (Boyd and Vandenberghe, 2004) (strong duality obtains for this problem because the Slater condition is satisfied by $p = \frac{1}{n}\mathbf{1}$) yields the maximization problem

$$\underset{\lambda \geq 0}{\text{maximize}} \; f(\lambda) := \inf_p \left\{ \frac{\lambda}{2}\left\|p - \frac{1}{n}\mathbf{1}\right\|_2^2 - \frac{\lambda\rho}{n} + p^\top z \mid p \geq 0, \; \mathbf{1}^\top p = 1 \right\}. \tag{43}$$

If we can efficiently compute the infimum (43), then it is possible to binary search over λ. Recall the standard fact (Hiriart-Urruty and Lemaréchal, 1993, Chapter VI.4.4) that for a collection $\{f_p\}_{p \in \mathcal{P}}$ of concave functions, if the infimum $f(x) = \inf_{p \in \mathcal{P}} f_p(x)$ 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 (43) becomes $f(\lambda) = \frac{\lambda}{2}\|p(\lambda) - \frac{1}{n}\mathbf{1}\|_2^2 - \frac{\lambda\rho}{n} + p(\lambda)^\top z$, whose derivative with respect to λ (holding p fixed) is $f'(\lambda) = \frac{1}{2}\|p(\lambda) - \frac{1}{n}\mathbf{1}\|_2^2 - \frac{\rho}{n}$.

Now we use well-known results on the Euclidean projection of a vector to the probability simplex (Duchi et al., 2008) to provide an efficient computation of the infimum (43). First, we assume with no loss of generality that $z_1 \le z_2 \le \cdots \le z_n$ and that $\mathbf{1}^\top z = 0$, because neither of these changes the original optimization problem (as $\mathbf{1}^\top p = 0$ 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

$$s_i = \sum_{j \le i} z_j, \quad \sigma_i^2 = \sum_{j \le i} z_j^2,$$

and we let z^2 be the vector whose entries are z_i^2. The infimum problem (43) is equivalent to projecting the vector $v(\lambda) \in \mathbb{R}^n$ defined by

$$v_i = \frac{1}{n} - \frac{1}{\lambda} z_i$$

onto the probability simplex. Notably (Duchi et al., 2008), the projection $p(\lambda)$ has the form $p_i(\lambda) = (v_i - \eta)_+$ for some $\eta \in \mathbb{R}$, where η is chosen such that $\sum_{i=1}^n p_i(\lambda) = 1$. Finding such a value η is equivalent (Duchi et al., 2008, Figure 1) to finding the unique index i such that

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

taking $i = n$ if no such index exists (the sum $\sum_{j=1}^i (v_j - v_i)$ is increasing in i and $v_1 - v_1 = 0$). Given the index i, algebraic manipulations show that $\eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{i} \sum_{j=1}^i z_j / \lambda = \frac{1}{n} - \frac{1}{i} - \frac{1}{i} s_i / \lambda$ satisfies the equality $\sum_{i=1}^n (v_i - \eta)_+ = 1$ and that $v_j - \eta \ge 0$ for all $j \le i$ while $v_j - \eta \le 0$ for $j > i$. Of course, given the index i and η, we may calculate the derivative $\frac{\partial}{\partial \lambda} f(\lambda)$ efficiently as well:

$$\begin{aligned}
f'(\lambda) &= \frac{\partial}{\partial \lambda} \left\{ \frac{\lambda}{2} \left\| p(\lambda) - n^{-1} \mathbf{1} \right\|_2^2 - \frac{\lambda \rho}{n} + p(\lambda)^\top z \right\} \\
&= \frac{1}{2} \left\| p(\lambda) - n^{-1} \mathbf{1} \right\|_2^2 - \frac{\rho}{n} = \frac{1}{2} \sum_{j=1}^{i} (v_j - \eta - n^{-1})^2 + \frac{1}{2} \sum_{j=i+1}^{n} \frac{1}{n^2} - \frac{\rho}{n} \\
&= \frac{1}{2} \sum_{j=1}^{i} \left(\frac{1}{\lambda} z_j + \eta \right)^2 + \frac{n-i}{2n^2} - \frac{\rho}{n} = \frac{\sigma_i^2}{2\lambda^2} + \frac{i \eta^2}{2} + \frac{s_i \eta}{\lambda} + \frac{n-i}{2n^2} - \frac{\rho}{n}.
\end{aligned}$$

Finding the index optimal i can be done by a binary search, which requires $O(\log n)$ time, and $f'(\lambda)$ is then computable in $O(1)$ time using the vectors s and σ^2. It is then possible to perform a binary search over λ using $f'(\lambda)$, which which requires $\log \frac{1}{\epsilon}$ iterations to find λ within accuracy ϵ, from which it is easy to compute $p(\lambda)$ via $p_i(\lambda) = (v_i - \eta)_+ = \left(n^{-1} - \lambda^{-1} z_i - \eta\right)_+$.

We summarize this discussion with pseudo-code in Figures 6 and 7, which provide a main routine and sub-routine for finding the optimal vector p. These routines show that, once provided the sorted vector z with $z_1 \le z_2 \le \cdots \le z_n$ (which requires $n \log n$ time to compute), we require only $O(\log \frac{1}{\epsilon} \cdot \log n)$ computations.

Inputs: Sorted vector $z \in \mathbb{R}^n$ with $\mathbf{1}^\top z = 0$, parameter $\rho > 0$, solution accuracy ϵ

Set $\lambda_{\min} = 0$ and $\lambda_{\max} = \lambda_\infty = \max\{n \|z\|_\infty, \sqrt{n/2\rho}\|z\|_2\}$
Set $s_i = \sum_{j \le i} z_j$ and $\sigma_i^2 = \sum_{j \le i} z_j^2$
While $|\lambda_{\max} - \lambda_{\min}| > \epsilon \lambda_\infty$
  Set $\lambda = \frac{\lambda_{\max} + \lambda_{\min}}{2}$
  Set $(\eta, i) = \textsc{FindShift}(z, \lambda, s)$ // (Figure 7)
  Set $f'(\lambda) = \frac{1}{2\lambda^2}\sigma_i^2 + \frac{\eta^2}{2} i^2 + \frac{\eta}{\lambda} s_i + \frac{n-i}{2n^2} - \frac{\rho}{n}$
  If $f'(\lambda) > 0$
    Set $\lambda_{\min} = \lambda$
  Else
    Set $\lambda_{\max} = \lambda$
Set $\lambda = \frac{1}{2}(\lambda_{\max} + \lambda_{\min})$, $(\eta, i) = \textsc{FindShift}(z, \lambda, s)$
Set $p_i = \left(\frac{1}{n} - \frac{1}{\lambda} z_i - \eta\right)_+$ and return p

Figure 6. Procedure FindP to find the vector p minimizing $\sum_{i=1}^n p_i z_i$ subject to the constraint $\frac{1}{2n}\|np - \mathbf{1}\|_2^2 \le \rho$. Method takes $\log \frac{1}{\epsilon}$ iterations of the loop.

Inputs: Sorted vector z with $\mathbf{1}^\top z = 0$, $\lambda > 0$, vector s with $s_i = \sum_{j \le i} z_j$

Set $i_{\text{low}} = 1, i_{\text{high}} = n$
If $\frac{1}{n} - \frac{z_n}{\lambda} \ge 0$
  Return $(\eta = 0, i = n)$
While $i_{\text{low}} \ne i_{\text{high}}$
  $i = \frac{1}{2}(i_{\text{low}} + i_{\text{high}})$
  $s_{\text{left}} = \frac{1}{\lambda}(i z_i - s_i)$ // (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 $s_{\text{right}} = \sum_{j=1}^{i+1} (v_j - v_{i+1})$)
  If $s_{\text{right}} \ge 1$ and $s_{\text{left}} < 1$
    Set $\eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{\lambda i} s_i$ and return (η, i)
  Else if $s_{\text{left}} \ge 1$
    Set $i_{\text{high}} = i - 1$
  Else
    Set $i_{\text{low}} = i + 1$
Set $i = i_{\text{low}}$ and $\eta = \frac{1}{n} - \frac{1}{i} - \frac{1}{\lambda i} s_i$ and return (η, i)

Figure 7. Procedure FindShift to find index i and parameter η such that, for the definition $v_i = \frac{1}{n} - \frac{1}{\lambda} z_i$, we have $v_j - \eta \ge 0$ for $j \le i$, $v_j - \eta \le 0$ for $j > i$, and $\sum_{j=1}^n (v_j - \eta)_+ = 1$. Method requires time $O(\log n)$.

References

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

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

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

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

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

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.

A. Ben-Tal, E. Hazan, T. Koren, and S. Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015.

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

M. Birman and M. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Sbornik: Mathematics*, 2(3):295–317, 1967.

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

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

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

O. Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, L'Ecole Polytechnique, 2002b.

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

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

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

S. Chatterjee. An error bound in the Sudakov-Fernique inequality. *arXiv:0510424 [math.PR]*, 2005.

N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

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

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

J.-y. Gotoh, M. J. Kim, and A. Lim. Robust empirical optimization is almost the same as mean-variance optimization. *Available at SSRN 2827400*, 2015.

C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

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

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

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

E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

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

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

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.

S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct):759–771, 2003.

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

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 30*, 2016.

A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1): 90–120, 1990.

A. B. Owen. *Empirical likelihood*. CRC press, 2001.

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

P. K. Shivaswamy and T. Jebara. Empirical Bernstein boosting. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

P. K. Shivaswamy and T. Jebara. Variance penalizing AdaBoost. In *Advances in Neural Information Processing Systems 25*, 2011.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *nips2010*, pages 2199–2207, 2010.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

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

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

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

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

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.

V. N. Vapnik and A. Y. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (In Russian).

D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.

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.

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.