Variance-based Regularization with Convex Objectives

Hongseok Namkoong
Stanford University
hnamk@stanford.edu

John C. Duchi
Stanford University
jduchi@stanford.edu

Abstract

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

1 Introduction

Let $\mathcal{X}$ be a sample space, P_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 [2, 6, 12, 7, 3] have shown results of the form that with high probability,

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

where C_1 and C_2 depend on the parameters of problem (1) and the desired confidence guarantee. Such bounds justify empirical risk minimization, 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.

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

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

where $\mathrm{Var}_{\widehat{P}_n}$ denotes the empirical variance. Maurer and Pontil [16] consider this idea, giving guarantees on the convergence and good performance of such a procedure. Unfortunately, even when

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

the loss ℓ is convex in θ, the formulation (3) is generally non-convex, which limits the applicability of procedures that minimize the variance-corrected empirical risk (3). In this paper, we develop an approach based on Owen's empirical likelihood [19] and ideas from distributionally robust optimization [4, 5, 10] that—whenever the loss ℓ is convex—provides a tractable *convex* formulation closely approximating the penalized risk (3). We give a number of theoretical guarantees and empirical evidence for its performance.

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

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

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

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

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 regardless of the value of $\rho \ge 0$ whenever the original loss $\ell(\cdot; X)$ is convex and Θ is a convex set. Namkoong and Duchi [18] propose a stochastic procedure for minimizing (4) almost as fast as stochastic gradient descent. See Appendix C for a detailed account of an alternative method.

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

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

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

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

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

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

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

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

with high probability. When there are parameters θ with small risk $R(\theta)$ (relative to the optimal parameter $\theta^\star$) and small variance $\mathrm{Var}(\ell(\theta, X))$, this guarantees that the excess risk $R(\widehat{\theta}_n^{\mathrm{rob}}) - R(\theta^\star)$ is essentially of order $O(\rho/n)$, where ρ governs our desired confidence level. We give an explicit example in Section 3.2 where our robustly regularized procedure (6) converges at $O(\log n/n)$ compared to $O(1/\sqrt{n})$ of empirical risk minimization.

Bounds that trade between risk and variance are known in a number of cases in the empirical risk minimization literature [15, 22, 2, 1, 6, 3, 7, 12], which is relevant when one wishes to achieve "fast

rates" of convergence for statistical learning algorithms. In many cases, such tradeoffs require either conditions such as the Mammen-Tsybakov noise condition [15, 6] or localization results [3, 2, 17] made possible by curvature conditions that relate the risk and variance. The robust solutions (6) enjoy a variance-risk tradeoff that is differen but holds essentially without conditions except compactness of Θ. We show in Section 3.3 that the robust solutions enjoy fast rates of convergence under typitcal curvature conditions on the risk R.

We complement our theoretical results in Section 4, where we conclude by providing two experiments comparing empirical risk minimization (ERM) strategies to robustly-regularized risk minimization (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.

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 (5). Although the variance of the loss is in general non-convex, the robust formulation (6) is a convex optimization problem for variance regularization whenever the loss function is convex [cf. 11, Prop. 2.1.2.].

To gain intuition for the variance expansion that follows, we consider the following equivalent formulation for the robust objective $\sup_{P \in \mathcal{P}_n} \mathbb{E}_P[Z]$

$$\operatorname*{maximize}_p \sum_{i=1}^n p_i z_i \;\; \text{subject to } p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}_+^n : \frac{1}{2} \left\| np - \mathbf{1} \right\|_2^2 \le \rho, \langle \mathbf{1}, p \rangle = 1 \right\}, \tag{7}$$

where $z \in \mathbb{R}^n$ is a vector. For simplicity, let $s_n^2 = \frac{1}{n} \left\| z \right\|_2^2 - (\overline{z})^2 = \frac{1}{n} \left\| z - \overline{z} \right\|_2^2$ denote the empirical "variance" of the vector z, where $\overline{z} = \frac{1}{n} \left\langle \mathbf{1}, z \right\rangle$ is the mean value of z. Then by introducing the variable $u = p - \frac{1}{n}\mathbf{1}$, the objective in problem (7) 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 (7) is equivalent to solving

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

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

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

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

$$\min_{i \in [n]} \frac{\sqrt{2\rho}(z_i - \overline{z})}{\sqrt{n s_n^2}} \ge -1. \tag{8}$$

Thus, if inequality (8) 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 (8), then, we have

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

A slight elaboration of this argument, coupled with the application of a few concentration inequalities, yields the next theorem. Recall 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 $[M_0, M_1]$ where $M = M_1 - M_0$ and fix $\rho \geq 0$. Then*

$$\left(\sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(Z)} - \frac{2M\rho}{n}\right)_+ \leq \sup_P \left\{\mathbb{E}_P[Z] : D_\phi(P\|\widehat{P}_n) \leq \frac{\rho}{n}\right\} - \mathbb{E}_{\widehat{P}_n}[Z] \leq \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(Z)}. \tag{9}$$

If $n \geq \max\{\frac{24\rho}{\mathrm{Var}(Z)}, \frac{16}{\mathrm{Var}(Z)}, 1\}M^2$ and we set $t_n = \sqrt{\mathrm{Var}(Z)}\left(\sqrt{1-n^{-1}} - \frac{1}{2}\right) - \frac{M^2}{n} \geq \sqrt{\frac{\mathrm{Var}(Z)}{18}}$,

$$\sup_{P: D_\phi(P\|\widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[Z] = \mathbb{E}_{\widehat{P}_n}[Z] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(Z)} \tag{10}$$

with probability at least $1 - \exp(-\frac{nt_n^2}{2M^2}) \geq 1 - \exp(-\frac{n\mathrm{Var}(Z)}{36M^2})$.

See Appendix A.1 for the proof. Inequality (9) and the exact expansion (10) 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.

We also provide a uniform variant of Theorem 1 based on the standard notion of the covering number, which we now define. Let $\mathcal{V}$ be a vector space with (semi)norm $\|\cdot\|$ on $\mathcal{V}$, and let $V \subset \mathcal{V}$. We say a collection $v_1, \ldots, v_N \subset V$ is an *ϵ-cover* of V if for each $v \in V$, there exists v_i such that $\|v - v_i\| \leq \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is then $N(V, \epsilon, \|\cdot\|) := \inf\{N \in \mathbb{N} : \text{ there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\|\}$. Now, let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, and define the $L^\infty(\mathcal{X})$-norm by $\|f - g\|_{L^\infty(\mathcal{X})} := \sup_{x\in\mathcal{X}} |f(x) - g(x)|$. Although we state our results abstractly, we typically take $\mathcal{F} := \{\ell(\theta, \cdot) \mid \theta \in \Theta\}$.

As a motivating example, we give the following standard bound on the covering number of Lipschitz losses [24].

Example 1: Let $\Theta \subset \mathbb{R}^d$ and assume that $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ is L-Lipschitz in θ with respect to the ℓ_2-norm for all $x \in \mathcal{X}$, meaning that $|\ell(\theta, x) - \ell(\theta', x)| \leq L\|\theta - \theta'\|_2$. 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 $\min_i |\ell(\theta, x) - \ell(\theta_i, x)| \leq L\epsilon$ for all θ, x. That is,

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

where $\mathrm{diam}(\Theta) = \sup_{\theta,\theta'\in\Theta} \|\theta - \theta'\|_2$. Thus ℓ_2-covering numbers of Θ control L^∞-covering numbers of the family $\mathcal{F}$. ♦

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 $f : \mathcal{X} \to [M_0, M_1]$ where $M = M_1 - M_0$, and let $\tau \geq 0$ be a constant. Define $\mathcal{F}_{\geq\tau} := \left\{f \in \mathcal{F} : \mathrm{Var}(f) \geq \tau^2\right\}$ and $t_n = \tau(\sqrt{1-n^{-1}} - \frac{1}{2}) - \frac{M^2}{n}$. If $\tau^2 \geq \frac{32\rho M^2}{n}$, then with probability at least $1 - N\left(\mathcal{F}, \frac{\tau}{32}, \|\cdot\|_{L^\infty(\mathcal{X})}\right)\exp\left(-\frac{nt_n^2}{2M^2}\right)$, we have for all $f \in \mathcal{F}_{\geq\tau}$*

$$\sup_{P: D_\phi(P\|\widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[f(X)] = \mathbb{E}_{\widehat{P}_n}[f(X)] + \sqrt{\frac{2\rho}{n}\mathrm{Var}_{\widehat{P}_n}(f(X))}. \tag{11}$$

We prove the theorem in Section A.2. Theorem 2 shows that the variance expansion of Theorem 1 holds uniformly for all functions f with sufficient variance. See Duchi, Glynn, and Namkoong [10] for an asymptotic analogue of the equality (11) for heavier tailed random variables.

3 Optimization by Minimizing the Robust Loss

Based on the variance expansions in the preceding section, we show that the robust solution (6) automatically trades between approximation and estimation error. In addition to $\|\cdot\|_{L^\infty(\mathcal{X})}$-covering

numbers defined in the previous section, we use the tighter notion of empirical ℓ_∞-covering numbers. For $x \in \mathcal{X}^n$, define $\mathcal{F}(x) = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ and 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 $N_\infty(\mathcal{F}) \le N(\mathcal{F})$.

Typically, we consider the function class $\mathcal{F} := \{\ell(\theta, \cdot) : \theta \in \Theta\}$, though we state our minimization results abstractly. Although the below result is in terms of covering numbers for ease of exposition, a variant holds depending on localized Rademacher averages [2] of the class $\mathcal{F}$, which can yield tighter guarantees (we omit such results for lack of space). We prove the following theorem in Section A.3.

Theorem 3. *Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to [M_0, M_1]$ with $M = M_1 - M_0$. Define the empirical minimizer*

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

Then for $\rho \ge t$, with probability at least $1 - 2(N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}) + 1)e^{-t}$,

$$\mathbb{E}[\widehat{f}(X)] \le \sup_{P : D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\widehat{f}(X)] + \frac{7M\rho}{n} + \left(2 + \sqrt{\frac{2t}{n-1}}\right)\epsilon \tag{12a}$$

$$\le \inf_{f \in \mathcal{F}} \left\{ \mathbb{E}[f] + 2\sqrt{\frac{2\rho}{n}\mathrm{Var}(f)} \right\} + \frac{11M\rho}{n} + \left(2 + \sqrt{\frac{2t}{n-1}}\right)\epsilon. \tag{12b}$$

Further, for $n \ge \frac{8M^2}{t}$, $t \ge \log 12$, and $\rho \ge 9t$, with probability at least $1 - 2(3N_\infty(\mathcal{F}, \epsilon, 2n) + 1)e^{-t}$,

$$\mathbb{E}[\widehat{f}(X)] \le \sup_{P : D_\phi(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\widehat{f}(X)] + \frac{11}{3}\frac{M\rho}{n} + \left(2 + 4\sqrt{\frac{2t}{n}}\right)\epsilon \tag{13a}$$

$$\le \inf_{f \in \mathcal{F}} \left\{ \mathbb{E}[f] + 2\sqrt{\frac{2\rho}{n}\mathrm{Var}(f)} \right\} + \frac{19M\rho}{3n} + \left(2 + 4\sqrt{\frac{2t}{n}}\right)\epsilon. \tag{13b}$$

Unlike analogous results for empirical risk minimization [6], Theorem 3 does not require the self-bounding type assumption $\mathrm{Var}(f) \le B\mathbb{E}[f]$. A consequence of this is that when $v = \mathrm{Var}(f^*)$ is small, where $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}[f]$, we achieve $O(1/n + \sqrt{v/n})$ (fast) rates of convergence. This condition is different from the typical conditions required for empirical risk minimization to have fast rates of convergence, highlighting the possibilities of variance-based regularization. It will be interesting to understand appropriate low-noise conditions (e.g. the Mammen-Tsybakov noise condition [15, 6]) guaranteeing good performance. Additionally, the robust objective $R_n(\theta, \mathcal{P}_n)$ is an empirical likelihood confidence bound on the population risk [10], and as empirical likelihood confidence bounds are self-normalizing [19], other fast-rate generalizations may exist.

3.1 Consequences of Theorem 3

We now 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. See Section A.4 for its proof.

Corollary 3.1. *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* (13) *hold with probability at least $1 - \left(c\,\mathsf{VC}(\mathcal{F})\left(\frac{16Mne}{\epsilon}\right)^{\mathsf{VC}(\mathcal{F})-1} + 2\right)e^{-t}$.*

Next, we focus more explicitly on the estimator $\widehat{\theta}_n^{\mathrm{rob}}$ 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 3.1, we have $\mathsf{VC}(\mathcal{F}) \le d$, and we obtain the following 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 3.1, we obtain the following result.

Corollary 3.2. *Let the conditions of the previous paragraph hold and assume that $\ell(\theta, x) \in [0, M]$ for all $\theta \in \Theta, x \in \mathcal{X}$. Then if $n \ge \rho \ge 9 \log 12$,*

$$R(\widehat{\theta}_n^{\rm rob}) \le R_n(\widehat{\theta}_n^{\rm rob}, \mathcal{P}_n) + \frac{11M\rho}{3n} + \frac{4M}{n} \le \inf_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho}{n} \mathrm{Var}(\ell(\theta; X))} \right\} + \frac{11M\rho}{n}$$

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

Unpacking Theorem 3 and Corollary 3.2 a bit, the first result (13a) provides a high-probability guarantees that the true expectation $\mathbb{E}[\widehat{f}]$ cannot be more than $O(1/n)$ worse than its robustly-regularized empirical counterpart, that is, $R(\widehat{\theta}_n^{\rm rob}) \le R_n(\widehat{\theta}_n^{\rm rob}, \mathcal{P}_n) + O(\rho/n)$, which is (roughly) a consequence of uniform variants of Bernstein's inequality. The second result (13b) guarantee the convergence of the empirical minimizer to a parameter with risk at most $O(1/n)$ larger than the best possible variance-corrected risk. In the case that the losses take values in $[0, M]$, then $\mathrm{Var}(\ell(\theta, X)) \le MR(\theta)$, and thus for $\epsilon = 1/n$ in Theorem 3, we obtain

$$R(\widehat{\theta}_n^{\rm rob}) \le R(\theta^\star) + C\sqrt{\frac{M\rho R(\theta^\star)}{n}} + C\frac{M\rho}{n},$$

a type of result well-known and achieved by empirical risk minimization for bounded nonnegative losses [6, 26, 25]. In some scenarios, however, the variance may satisfy $\mathrm{Var}(\ell(\theta, X)) \ll MR(\theta)$, yielding improvements.

To give an alternative variant of Corollary 3.2, let $\Theta \subset \mathbb{R}^d$ and 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 $D := \mathrm{diam}(\Theta) = \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\| < \infty$, then $0 \le \ell(\theta, x) \le L \,\mathrm{diam}(\Theta) =: M$.

Corollary 3.3. *Let the conditions of the preceeding paragraph hold. Set $t = \rho = \log 2n + d \log(2nDL)$ and $\epsilon = \frac{1}{n}$ in Theorem 3 and assume that $D \lesssim n^k$ and $L \lesssim n^k$ for a numerical constant k. With probability at least $1 - 1/n$,*

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

where C is a numerical constant.

3.2 Beating empirical risk minimization

We now provide an example in which the robustly-regularized estimator (6) exhibits a substantial improvement over empirical risk minimization. We expect the robust approach to offer performance benefits in situations in which the empirical risk minimizer is highly sensitive to noise, say, because the losses are piecewise linear, and slight under- or over-estimates of slope may significantly degrade solution quality. With this in mind, we construct a toy 1-dimensional example—estimating the median of a distribution supported on $\mathcal{X} = \{-1, 0, 1\}$—in which the robust-regularized estimator has convergence rate $\log n/n$, while empirical risk minimization is at best $1/\sqrt{n}$.

Define the loss $\ell(\theta; x) = |\theta - x| - |x|$, and for $\delta \in (0, 1)$ let the distribution P be defined by $P(X = 1) = \frac{1-\delta}{2}$, $P(X = -1) = \frac{1-\delta}{2}$, $P(X = 0) = \delta$. Then for $\theta \in \mathbb{R}$, the risk of the loss is

$$R(\theta) = \delta|\theta| + \frac{1-\delta}{2}|\theta - 1| + \frac{1-\delta}{2}|\theta + 1| - (1 - \delta).$$

By symmetry, it is clear that $\theta^\star := \mathrm{argmin}_\theta \, R(\theta) = 0$, which satisfies $R(\theta^\star) = 0$. (Note that $\ell(\theta, x) = \ell(\theta, x) - \ell(\theta^\star, x)$.) Without loss of generality, we assume that $\Theta = [-1, 1]$. Define the empirical risk minimizer and the robust solution

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

Intuitively, if too many of the observations satisfy $X_i = 1$ or too many satisfy $X_i = -1$, then $\widehat{\theta}^{\rm erm}$ will be either 1 or -1; for small δ, such events become reasonably probable. On the other hand, we have $\ell(\theta^\star; x) = 0$ for all $x \in \mathcal{X}$, so that $\mathrm{Var}(\ell(\theta^\star; X)) = 0$ and variance regularization achieves the rate $O(\log n/n)$ as opposed to empirical risk minimizer's $O(1/\sqrt{n})$. See Section A.6 for the proof.

Proposition 1. *Under the conditions of the previous paragraph, for $n \geq \rho = 3 \log n$, with probability at least $1 - \frac{4}{n}$, we have $R(\widehat{\theta}_n^{\rm rob}) - R(\theta^\star) \leq \frac{45 \log n}{n}$. However, with probability at least $2\Phi(-\sqrt{\frac{n}{n-1}}) - 2\sqrt{2}/\sqrt{\pi e n} \geq 2\Phi(-\sqrt{\frac{n}{n-1}}) - n^{-\frac{1}{2}}$, we have $R(\widehat{\theta}^{\rm erm}) \geq R(\theta^\star) + n^{-\frac{1}{2}}$.*

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

3.3 Fast Rates

In cases in which the risk R has curvature, empirical risk minimization often enjoys faster rates of convergence [6, 21]. 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 $S^\epsilon := \{\theta \in \Theta : R(\theta) \leq \inf_{\theta^\star \in \Theta} R(\theta^\star) + \epsilon\}$ denote the ϵ-sub-optimal (solution) set, and similarly let $\widehat{S}^\epsilon := \{\theta \in \Theta : R_n(\theta, \mathcal{P}_n) \leq \inf_{\theta' \in \Theta} R_n(\theta', \mathcal{P}_n) + \epsilon\}$. For a vector $\theta \in \Theta$, let $\pi_S(\theta) = \operatorname{argmin}_{\theta^\star \in S} \|\theta^\star - \theta\|_2$ denote the Euclidean projection of θ onto the set S.

Our below result depends on a local notion of Rademacher complexity. For i.i.d. random signs $\varepsilon_i \in \{\pm 1\}$, the empirical Rademacher complexity of a function class $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ is

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

Although we state our results abstractly, we typically take $\mathcal{F} := \{\ell(\theta, \cdot) \mid \theta \in \Theta\}$. For example, when $\mathcal{F}$ is a VC-class, we typically have $\mathbb{E}[\mathfrak{R}_n \mathcal{F}] \lesssim \sqrt{\mathsf{VC}(\mathcal{F})/n}$. Many other bounds on $\mathbb{E}[\mathfrak{R}_n \mathcal{F}]$ are possible [1, 24, Ch. 2]. For $A \subset \Theta$ let $\mathfrak{R}_n(A)$ denote the Rademacher complexity of the localized process $\{x \mapsto \ell(\theta; x) - \ell(\pi_S(\theta); x) : \theta \in A\}$. We then have the following result, whose proof we provide in Section A.7.

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

$$R(\theta) - \inf_{\theta \in \Theta} R(\theta) \geq \lambda \operatorname{dist}(\theta, S)^\gamma \text{ for all } \theta \text{ such that } \operatorname{dist}(\theta, S) \leq r. \tag{14}$$

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

$$\epsilon \geq \left(\frac{8L^2\rho}{n}\right)^{\frac{\gamma}{2(\gamma-1)}} \left(\frac{2}{\lambda}\right)^{\frac{1}{\gamma-1}} \text{ and } \frac{\epsilon}{2} \geq 2\mathbb{E}[\mathfrak{R}_n(S^{2\epsilon})] + L\left(\frac{2\epsilon}{\lambda}\right)^{\frac{1}{\gamma}} \sqrt{\frac{2t}{n}}, \tag{15}$$

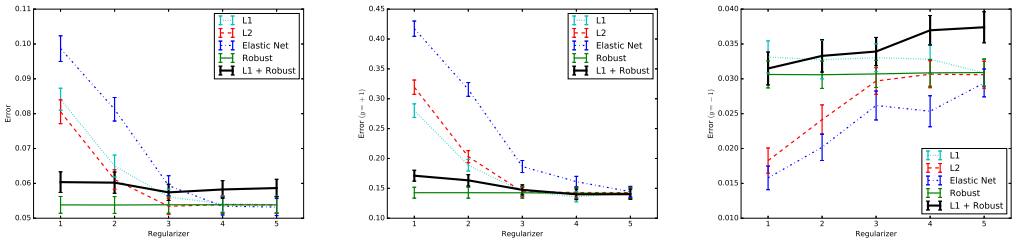
then $\mathbb{P}(\widehat{S}^\epsilon \subset S^{2\epsilon}) \geq 1 - e^{-t}$, and inequality (15) holds for all $\epsilon \gtrsim (\frac{L^2(t+\rho+d)}{\lambda^{2/\gamma} n})^{\frac{\gamma}{2(\gamma-1)}}$.

4 Experiments

We present two real classification experiments to carefully compare standard empirical risk minimization (ERM) to the variance-regularized approach we present. Empirically, we show that the ERM estimator $\widehat{\theta}^{\rm erm}$ performs poorly on rare classes with (relatively) more variance, where the robust solution achieves improved classification performance on rare instances. In all our experiments, this occurs with little expense over the more common instances.

4.1 Protease cleavage experiments

For our first experiment, we compare our robust regularization procedure to other regularizers using the HIV-1 protease cleavage dataset from the UCI ML-repository [14]. 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).



(a) test error (b) rare class ($Y_i = +1$) (c) common class ($Y_i = -1$)

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

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 $\Theta = \left\{\theta \in \mathbb{R}^d : a_1 \left\|\theta\right\|_1 + a_2 \left\|\theta\right\|_2 \le r\right\}$, which is equivalent to elastic net regularization [27], 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 $r \in \{10^2, 2 \cdot 10^2, 10^3, 2 \cdot 10^3, 10^4\}$, our robust regularizer with $\rho \in \{10^2, 10^3, 10^4, 5 \cdot 10^4, 10^5\}$ and our robust regularizer coupled with the 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 zero-one (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 $1/10$ fraction (test).

We plot results summarizing these experiments in Figure 1. The horizontal axis in each figure indexes our choice of regularization value (so "`Regularizer = 1`" for the 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. Notably, even when coupled with a fairly stringent ℓ_1-constraint ($r = 100$), robust regularization has performance better than ℓ_1 except for large values r, especially on the rare label $Y = +1$.

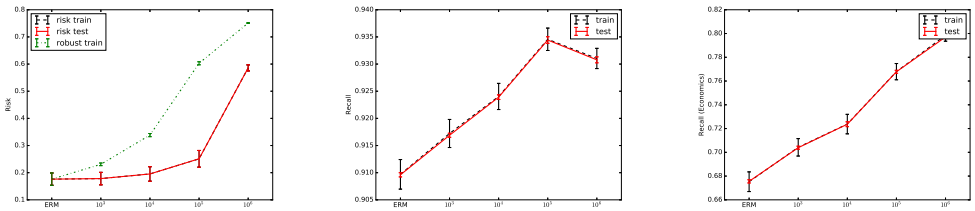
We investigate the effects of the robust regularizer with a slightly different perspective in Table 1, where we use $\Theta = \{\theta : \left\|\theta\right\|_1 \le 100\}$ for the constraint set for each experiment. We give error rates and logistic risk values for the different procedures, averaged over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

Table 1: 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

4.2 Document classification in the Reuters corpus

For our second experiment, we consider a multi-label classification problem with a reasonably large dataset. The Reuters RCV1 Corpus [13] has 804,414 examples with $d = 47{,}236$ features, where feature j is an indicator variable for whether word j appears in a given document. The goal is to classify documents as a subset of the 4 categories where documents are labeled with a subset of those. As documents can belong to multiple categories, we fit binary classifiers on each of the four



(a) (b) (c)

Figure 2: Reuters corpus experiment. (a) Logistic risks. (b) Recall. (c) Recall on Economics (rare).

categories. Each category has different number of documents (Corporate: $381,327$, Economics: $119,920$, Government: $239,267$, Markets: $204,820$) In this experiment, we expect the robust solution to outperform ERM on the rarer category (Economics), as the robustification (6) naturally upweights rarer (harder) instances, which disproportionally affect variance—as in the previous experiment.

For each category $k \in \{1,2,3,4\}$, we use the logistic loss $\ell(\theta_k;(x,y)) = \log(1+\exp(-y\theta_k^\top x))$. For each binary classifier, we use the ℓ_1 constraint set $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\right\}$. 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). We partition the data into ten equally-sized sub-samples and perform ten validation experiments, where in each experiment we use one of the ten subsets for fitting the logistic models and the remaining nine partitions as a test set to evaluate performance.

In Figure 2, we summarize the results of our experiment averaged over the 10 runs, with 2-standard error bars (computed across the folds). To facilitate comparison across the document categories, we give exact values of these averages in Tables 2 and 3. Both $\widehat{\theta}_n^{\mathrm{rob}}$ and $\widehat{\theta}^{\mathrm{erm}}$ have reasonably high precision across all categories, with increasing ρ giving a mild improvement in precision (from $.93 \pm .005$ to $.94 \pm .005$). On the other hand, we observe in Figure 2(c) that ERM has low recall ($.69$ on test) for the Economics category, which contains about 15% of documents. As we increase 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. This precision and recall improvement comes in spite of the increase in the average binary logistic risk for each of the 4 classes. In Figure 2(a), we plot the average binary logistic 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 error rate.

Table 2: 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 3: 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

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

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 Samsung Fellowship. JCD was also partially supported by the National Science Foundation award NSF-CAREER-1553086 and the Sloan Foundation.

References

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

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

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

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

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

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

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

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

[9] 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.

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

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

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

[13] 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.

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

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

[16] 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.

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

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

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

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

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

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

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

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

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

[26] 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.

[27] 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.

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

A Proofs of Main Results

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

A.1 Proof of Theorem 1

Let $\sigma^2 = \mathrm{Var}(Z)$ and $s_n^2 = \mathrm{Var}_{\widehat{P}_n}(Z) = \mathbb{E}_{\widehat{P}_n}[Z^2] - \mathbb{E}_{\widehat{P}_n}[Z]^2$ denote the population and sample variance of Z, respectively. 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 $\sigma^2, s_n^2 > 0$. Recall the maximization problem (7), 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 (8), which guarantees that the maximizing value of problem (7) 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 (8) it is certainly sufficient that

$$2\rho \frac{M^2}{n s_n^2} \le 1, \ \text{ or } \ n \ge \frac{2\rho M^2}{s_n^2}, \ \text{ or } \ s_n^2 \ge \frac{2\rho M^2}{n}. \tag{16}$$

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 (16) for the exact expansion to hold yields the two-sided variance bounds (9).

We now turn to showing the high-probability exact expansion (10), which occurs whenever the sample variance is large enough by expression (16). To that end, we show that s_n^2 is bounded from below with high probability. Define the event

$$\mathcal{E}_n := \left\{ s_n^2 \ge \frac{1}{4} \sigma^2 \right\},$$

and let $n \ge \max\left\{ \frac{16\rho}{\sigma^2}, \frac{16}{\sigma^2}, 1 \right\} M^2$. Then, we have on event $\mathcal{E}_n$

$$n \ge \frac{16\rho M^2}{\sigma^2} \ge \frac{16}{4} \frac{\rho M^2}{s_n^2} = \frac{4\rho M^2}{s_n^2},$$

so that the sufficient condition (16) holds and expression (10) follows. We now argue that the event $\mathcal{E}_n$ has high probability via the following lemma, which is essentially an application of known concentration inequalities for convex functions coupled with a few careful estimates of the expectation of standard deviations.

Lemma A.1. *Let Z_i be independent random variables taking values in $[M_0, M_1]$ with $M = M_1 - M_0$, and let $s_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i^2 - \left(\frac{1}{n}\sum_{i=1}^n Z_i\right)^2$. For all $t \ge 0$, we have*

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

See Section B.2 for a proof of the lemma. When the Z_i are i.i.d., we obtain

$$\mathbb{P}\left(s_n \ge \sigma\sqrt{1 - n^{-1}} + t \right) \ \vee \ \mathbb{P}\left(s_n \le \sigma\sqrt{1 - n^{-1}}\sigma - \frac{M^2}{n} - t \right) \le \exp\left(-\frac{nt^2}{2M^2} \right)$$

where $\sigma^2 = \mathrm{Var}(Z)$.

Now, substitute $t = \sigma(\sqrt{1-n^{-1}} - \frac{1}{2}) - \frac{M^2}{n}$ so that

$$\sigma(1 - n^{-\frac{1}{2}}) - \frac{M^2}{n} - t = \frac{1}{2}\sigma.$$

Note that $\frac{M^2}{n} \le \frac{M}{\sqrt{n}} \le \sigma/4$ and since $\sqrt{1-n^{-1}} \ge 1 - \frac{1}{2n\sqrt{1-n^{-1}}}$ and $\sigma^2 \le M^2/4$ by standard variance bounds, our choice of n also satisfies $n \ge 16M^2/\sigma^2 \ge 64$. We thus have $t/\sigma \ge 1 - \frac{1}{16\sqrt{63}} - \frac{1}{2} - \frac{1}{4} > \frac{1}{\sqrt{18}}$. We obtain

$$\mathbb{P}\left(\mathcal{E}_n\right) \ge 1 - \exp\left(-\frac{n\sigma^2\left((1-n^{-1})^{1/2} - 1/2 - M^2/\sigma n\right)_+^2}{2M^2}\right) \ge 1 - \exp\left(-\frac{n\sigma^2}{36M^2}\right).$$

This gives the result (10).

A.2 Proof of Theorem 2

Throughout this proof, we let $s_n^2(f) = \mathbb{E}_{\widehat{P}_n}[f(X)^2] - \mathbb{E}_{\widehat{P}_n}[f(X)]^2$ denote the empirical variance of the function f, and we use $\sigma_Q^2(f) = \mathbb{E}_Q[(f - \mathbb{E}_Q[f])^2]$ to denote the variance of f under the distribution Q. Our starting point is to recall from inequality (16) in the proof of Theorem 1 that for each $f \in \mathcal{F}$, the empirical variance equality (11) holds if $n \ge \frac{4\rho M^2}{s_n^2(f)}$. As a consequence, Theorem 2 will follow if we can provide a uniform lower bound on the sample variances $s_n^2(f)$ that holds with high enough probability.

Set $\epsilon > 0$, and let $\{f_1, \ldots, f_N\}$, where $N = N(\mathcal{F}_{\ge\tau}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})})$, be a minimal ϵ-cover of $\mathcal{F}_{\ge\tau}$. Define the event

$$\mathcal{E}_n := \left\{ s_n^2(f_i) \ge \frac{1}{4}\sigma^2(f_i) \text{ for } i = 1, \ldots, N \right\}.$$

By applying Lemma A.1 and the argument immediately following it in the proof of Theorem 1, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_n) &\ge 1 - \sum_{i=1}^N \exp\left(-\frac{n\sigma^2(f_i)\left(1/2 - 1/\sqrt{n} - M^2/\sigma(f_j)n\right)_+^2}{2M^2}\right) \\
&\ge 1 - N(\mathcal{F}_{\ge\tau}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}) \exp\left(-\frac{n\tau^2\left(1/2 - 1/\sqrt{n} - M^2/\tau n\right)_+^2}{2M^2}\right). \qquad (17)
\end{aligned}$$

Thus, to obtain the theorem, we must show that the event $\mathcal{E}_n$ implies that the variance expansion (11) holds for each $g \in \mathcal{F}_{\ge\tau}$.

For $g \in \mathcal{F}$ and f_j such that $\|g - f_j\|_{L^\infty(\mathcal{X})} \le \epsilon$, the triangle inequality implies

$$\begin{aligned}
|\sigma_Q(g) - \sigma_Q(f_j)| &\le \sqrt{\mathbb{E}_Q[(f_j - g + \mathbb{E}_Q g - \mathbb{E}_Q f_j)^2]} \\
&= \sqrt{\mathbb{E}_Q[(f_j - g)^2] - (\mathbb{E}_Q[f_j - g])^2} \le \epsilon
\end{aligned}$$

for any distribution Q on $\mathcal{X}$. Then on the event $\mathcal{E}_n$, for any $g \in \mathcal{F}$ and the f_j closest to g from the covering, we have

$$s_n(g) \ge s_n(f_j) - \epsilon \ge \frac{1}{2}\sigma(f_j) - \epsilon \ge \frac{1}{2}\sigma(g) - \frac{3}{2}\epsilon.$$

That is, $s_n(g) \ge \frac{1}{2}\sigma(g) - \frac{3}{2}\epsilon$. Now, let $\epsilon = \frac{\tau}{24}$, which gives that

$$\mathcal{E}_n \text{ implies } \; s_n(g) \ge \frac{7}{16}\sigma(g) \ge \frac{7}{16}\tau \;\text{ for all } g \in \mathcal{F}_{\ge\tau}.$$

Recalling the sufficient condition (16) for the exact variance expansion to hold, we see that on $\mathcal{E}_n$,

$$\frac{4 \cdot 256\rho M^2}{49\tau^2} \ge \frac{4\rho M^2}{s_n^2(g)} \;\text{ for all } g \in \mathcal{F}_{\ge\tau}.$$

Taking $n \ge \frac{32\rho M^2}{\tau^2} > \frac{256\rho M^2}{49\tau^2}$ thus gives the result.

A.3 Proof of Theorem 3

Before proving the theorem proper, we state two technical lemmas that provide uniform Bernstein-like bounds for the class $\mathcal{F}$. The first applies for empirical ℓ_∞-covering numbers.

Lemma A.2 (Maurer and Pontil [16], Theorem 6)**.** *Let $n \geq \frac{8M^2}{t}$ and $t \geq \log 12$. Then with probability at least $1 - 6N_\infty(\mathcal{F}, \epsilon, 2n)e^{-t}$, we have*

$$\mathbb{E}[f] \leq \mathbb{E}_{\widehat{P}_n}[f] + 3\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f)t}{n}} + \frac{15Mt}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \tag{18}$$

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

The second lemma applies when we have uniform $\|\cdot\|_{L^\infty(\mathcal{X})}$-covering numbers for $\mathcal{F}$.

Lemma A.3. *Let $\mathcal{F}$ be a collection functions with covering numbers $N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})})$, and assume that $|f(x)| \leq M$ for all x. Then with probability at least $1 - 2N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})})e^{-t}$,*

$$\mathbb{E}[f] \leq \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f)t}{n-1}} + \frac{\sqrt{2t}M^2}{n^{3/2} - n} + \frac{2 + 3\sqrt{2}}{3}\frac{Mt}{n} + \left(2 + \sqrt{\frac{2t}{n-1}}\right)\epsilon$$

and

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

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

As this lemma is essentially standard, we defer its proof to Section B.3.

We prove only the first set of bounds (13) in the theorem, which are based on Lemma A.2, as the proof of the bounds (12) follows in precisely the same way from Lemma A.3. We now return to the proof of Theorem 3. Let $\mathcal{E}_1$ denote that the event that the inequalities (18) hold. Now, let

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

Then because $\mathcal{E}_1$ holds, we have

$$\begin{aligned}
\mathbb{E}_P[\widehat{f}] &\leq \mathbb{E}_{\widehat{P}_n}[\widehat{f}] + \sqrt{\frac{18\mathrm{Var}_{\widehat{P}_n}(\widehat{f}(X))t}{n}} + \frac{15Mt}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \\
&\stackrel{(i)}{\leq} \sup_{P : D_\phi(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[\widehat{f}(X)] + \sqrt{\frac{2\rho\mathrm{Var}_{\widehat{P}_n}(\widehat{f}(X))}{n}} \\
&\quad - \left(\sqrt{\frac{2\rho\mathrm{Var}_{\widehat{P}_n}(\widehat{f}(X))}{n}} - \frac{2M\rho}{n}\right)_+ + \frac{5M\rho}{3n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \\
&\stackrel{(ii)}{\leq} \sup_{P : D_\phi(P \| \widehat{P}_n) \leq \frac{\rho}{n}} \mathbb{E}_P[f(X)] + \frac{11}{3}\frac{M\rho}{n} + 2\left(1 + 2\sqrt{\frac{2t}{n}}\right)\epsilon \text{ for all } f \in \mathcal{F},
\end{aligned} \tag{19}$$

where inequality (i) follows from the bounds (9) in Theorem 1 and the fact that $\rho \geq 9t$ by assumption and inequality (ii) because $\widehat{f}$ minimizes $\sup_{P : D_\phi(P \| \widehat{P}_n) \leq \rho/n} \mathbb{E}_P[f(X)]$. This gives the first result (13a).

For the second result (13b), we bound the supremum term in expression (19). As the function f is fixed, we have

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

with probability at least $1 - e^{-t}$, and we similarly have by Lemma A.1 that

$$\sqrt{\mathrm{Var}_{\widehat{P}_n}(f)} \le \sqrt{1 - n^{-1}}\sqrt{\mathrm{Var}(f)} + \sqrt{\frac{2tM^2}{n}}$$

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 (9) of Theorem 1 and inequality (ii) from our assumption that $\rho \ge t$. Substituting this expression into our earlier bound (19) 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.

A.4 Proof of Corollary 3.1

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 [24], we have

$$\sup_Q N(\mathcal{F}, \epsilon, \|\cdot\|_{L^1(Q)}) \le c\mathsf{VC}(\mathcal{F})\left(\frac{8Me}{\epsilon}\right)^{\mathsf{VC}(\mathcal{F})-1} \tag{20}$$

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 follows by applying the bound (20) for $\epsilon/2n$.

A.5 Proof of Corollary 3.3

.

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,

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

where $\mathrm{diam}(\Theta) = \sup_{\theta,\theta' \in \Theta} \|\theta - \theta'\|_2$. Thus ℓ_2-covering numbers of Θ control L^∞-covering numbers of the family $\mathcal{F}$. Plugging in the respective values of ρ, t and ϵ in Theorem 3, we obtain the result.

A.6 Proof of Proposition 1

We certainly have $\ell(\theta^\star; x) = 0$ for all $x \in \mathcal{X}$, so that $\mathrm{Var}(\ell(\theta^\star; X)) = 0$. Now, consider the bound in Theorem 3 (12b). 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 $\widehat{\theta}_n^{\mathrm{rob}} \in \mathrm{argmin}_{\theta\in\Theta} R_n(\theta, \mathcal{P}_n)$, then

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

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

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

We now show the claim for th empirical risk minimizer. Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$ denotes the standard Gaussian CDF. (See Section B.1 for a proof.)

Lemma A.4. *Let the loss $\ell(\theta; x) = |\theta - x| - |x|$, $\delta \in [0, 1]$, and X follow the distribution P given by $P(X = 1) = \frac{1-\delta}{2}$, $P(X = -1) = \frac{1-\delta}{2}$, $P(X = 0) = \delta$.. Then with probability at least*

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

we have $R(\widehat{\theta}^{\rm erm}) - R(\theta^\star) \ge \delta$.

The risk for the empirical risk minimizer, as Lemma A.4 shows, may be substantially higher; taking $\delta = 1/\sqrt{n}$ we see that with probability at least $2\Phi(-\sqrt{\frac{n}{n-1}}) - 2\sqrt{2}/\sqrt{\pi e n} \ge 2\Phi(-\sqrt{\frac{n}{n-1}}) - n^{-\frac{1}{2}}$,

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

A.7 Proof of Theorem 4

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

$$\Delta_n(\theta) := \mathbb{E}\left[\ell(\theta; X) - \ell(\pi(\theta); X)\right] - \mathbb{E}_{\widehat{P}_n}\left[\ell(\theta; X) - \ell(\pi(\theta); X)\right]. \tag{21}$$

We begin with the following

Claim A.1. *If $\widehat{S}^\epsilon \not\subset S^{2\epsilon}$, then*

$$\sup_{\theta \in S^{2\epsilon}} \left\{ \Delta_n(\theta) + \sqrt{\frac{2\rho}{n} \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X) - \ell(\pi(\theta); X))} \right\} \ge \epsilon. \tag{22}$$

Deferring the proof of the claim, let us prove the theorem. First, the growth condition (14) shows that

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

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

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

and so by the assumption (15) 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 (22) holds then

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

and recalling the definition (21) of Δ_n, it then follows that

$$\mathbb{P}\left(\widehat{S}^\epsilon \not\subset S^{2\epsilon}\right) \le \mathbb{P}\left(\sup_{\theta \in S^{2\epsilon}} \Delta_n(\theta) \ge \frac{\epsilon}{2}\right). \tag{23}$$

To bound the probability (23), we use standard bounded difference and symmetrization arguments [e.g. 7, Theorem 6.5]. Letting $f(X_1, \ldots, X_n) := \sup_{\theta \in S^{2\epsilon}} \Delta_n(\theta)$, the function f satisfies bounded differences:

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

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

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

for all $t \geq 0$. Letting $u = \frac{nt^2}{2L^2} \left(\frac{\lambda}{2\epsilon} \right)^{\frac{2}{\gamma}}$ above and recalling the assumption (15) upper bounding $\mathbb{E}[\mathfrak{R}_n(S^{2\epsilon})]$, we have $\mathbb{P}(\sup_{\theta \in S^{2\epsilon}} \Delta_n(\theta) \geq \frac{\epsilon}{2}) \leq e^{-u}$. The first result of the theorem follows from the bound (23).

To show the last claim of Theorem 4, note that a minor extension of standard chaining arguments (see, for example, van der Vaart and Wellner [24, Section 2.2]), we have

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

for some numerical constant $C > 0$. Plugging this into the bound (15) and rearranging for ϵ, we obtain the result.

Proof of Claim A.1 If $\widehat{S}^\epsilon \not\subset S^{2\epsilon}$, then certainly it is the case that there is some $\theta \in \Theta \setminus S^{2\epsilon}$ such that

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

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

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

For all $t \in [0, 1]$, we have by definition of orthogonal projection (because the vector $\theta - \pi(\theta)$ belongs to the normal cone to S and $\pi(\theta)$; cf. [11, Prop. III.5.3.3]) that $\pi(t\theta + (1-t)\pi(\theta)) = \pi(\theta)$. Thus, choosing t appropriately there exists $\theta' \in \operatorname{bd} S^{2\epsilon}$ with $\theta' = t\theta + (1-t)\pi(\theta)$ and $\pi(\theta') = \pi(\theta)$, and $R_n(\theta', \mathcal{P}_n) \leq R_n(\pi(\theta'), \mathcal{P}_n) + \epsilon$.

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

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

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

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

Applying the upper bound in inequality (9) of Theorem 1 gives the claim. □

B Proofs of Technical Lemmas

B.1 Proof of Lemma A.4

Defining $N_y := \operatorname{card}\{i \in [n] : X_i = y\}$ 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

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

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

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

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 [28] 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 [23, 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

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

B.2 Proof of Lemma A.1

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

Lemma B.1 (Samson [20], Corollary 3)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-Lipschitz continuous with respect to the ℓ_2-norm over $[a, b]^n$, and let $Z_1, \ldots, Z_n$ be independent random variables on $[a, b]$. Then for all $t \geq 0$,*

$$\mathbb{P}(f(Z_{1:n}) \geq \mathbb{E}[f(Z_{1:n})] + t) \vee \mathbb{P}(f(Z_{1:n}) \leq \mathbb{E}[f(Z_{1:n})] - t) \leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Lemma B.2. *Let Y_i be independent random variables with finite 4th moment. Then we have the following inequalities:*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] \geq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} - \frac{1}{\sqrt{n}} \sqrt{\frac{\frac{1}{n}\sum_{i=1}^n \operatorname{Var}(Y_i^2)}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}} \tag{24a}$$

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] \geq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} - \frac{1}{n} \frac{\frac{1}{n}\sum_{i=1}^n \operatorname{Var}(Y_i^2)}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}, \tag{24b}$$

and if $|Y_i| \leq C$ for with probability 1 for all $1 \leq i \leq n$, then

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] \geq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} - \min\left\{\frac{C^2}{n}, \frac{C}{\sqrt{n}}\right\}. \tag{24c}$$

We defer the proof of Lemma B.2 to the end of this section in Section B.2.1.

The function $\mathbb{R}^n \ni z \mapsto \left\|(I - (1/n)\mathbf{1}\mathbf{1}^\top)z\right\|_2$ is 1-Lipschitz with respect to the Euclidean norm, so Lemma B.1 implies

$$\mathbb{P}\left(\sqrt{\mathrm{Var}_{\widehat{P}_n}(Z)} \geq \mathbb{E}[\sqrt{\mathrm{Var}_{\widehat{P}_n}(Z)}] + t\right) \vee \mathbb{P}\left(\sqrt{\mathrm{Var}_{\widehat{P}_n}(Z)} \leq \mathbb{E}[\sqrt{\mathrm{Var}_{\widehat{P}_n}(Z)}] - t\right) \leq \exp\left(-\frac{nt^2}{2M^2}\right).$$

Then

$$\mathbb{E}[\sqrt{\mathrm{Var}_{\widehat{P}_n}(Z)}] \leq \sqrt{\mathbb{E}[\mathrm{Var}_{\widehat{P}_n}(Z)]} = \sqrt{1 - n^{-1}}\sqrt{\mathrm{Var}(Z)},$$

giving the first inequality of the lemma. For the second, let $Y_i = Z_i - \frac{1}{n}\sum_{j=1}^n Z_j$ so that $s_n^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2$. Applying Lemma B.2 with $C = M$, we obtain $\mathbb{E}[s_n] \geq \sqrt{\mathbb{E}[s_n^2]} - \frac{M^2}{n}$ which gives the result.

B.2.1 Proof of Lemma B.2

We first prove the claim (24a). To see this, we use that

$$\inf_{\lambda \geq 0}\left\{\frac{a^2}{2\lambda} + \frac{\lambda}{2}\right\} = \sqrt{a^2} = |a|,$$

and taking derivatives yields that for all $\lambda' \geq 0$,

$$\frac{a^2}{2\lambda} + \frac{\lambda}{2} \geq \frac{a^2}{2\lambda'} + \frac{\lambda'}{2} - \left(\frac{a^2}{2\lambda'^2} - \frac{1}{2}\right)(\lambda - \lambda').$$

By setting $\lambda_n = \sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}$, we thus have for any $\lambda \geq 0$ that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i^2}{2n\lambda_n} + \frac{\lambda_n}{2}\right] \\ &\geq \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i^2}{2n\lambda} + \frac{\lambda}{2}\right] + \mathbb{E}\left[\left(\frac{1}{2} - \frac{\sum_{i=1}^n Y_i^2}{2n\lambda^2}\right)(\lambda_n - \lambda)\right].\end{aligned}$$

Now we take $\lambda = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}$, and we apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] \\ &\geq \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} - \frac{1}{2\lambda^2}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (Y_i^2 - \mathbb{E}[Y_i^2])\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}}\right)^2\right]^{\frac{1}{2}} \\ &= \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} - \frac{1}{2\sqrt{n}\lambda^2}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Y_i^2)\right)^{\frac{1}{2}} \mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}}\right)^2\right]^{\frac{1}{2}}\end{aligned}$$

where the last equality follows from independence. Using the triangle inequality, we obtain that the final expectation is bounded by $2\lambda = 2\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}$, which gives inequality (24a). Now we give a sharper result. We have

$$\begin{aligned}&\mathbb{E}\left[\left(\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}}\right)^2\right] \\ &= \frac{2}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2] - 2\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]\right)^{\frac{1}{2}} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{\frac{1}{2}}\right] \\ &\leq \frac{2}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Y_i^2)\right)^{\frac{1}{2}}\end{aligned}$$

where we used the first inequality (24a). Thus we also obtain the lower bound (24b). The final inequality follows immediately upon noticing that $\mathrm{Var}(Y^2) \le \mathbb{E}[Y^4] \le C^2\mathbb{E}[Y^2]$ for $\|Y\|_\infty \le C$.

B.3 Proof of Lemma A.3

By the standard Bernstein inequalities, we have that for any fixed f,

$$\mathbb{E}_{\widehat{P}_n}[f] \le \mathbb{E}[f] + \sqrt{\frac{2\mathrm{Var}(f)t}{n}} + \frac{2M}{3n}t \text{ with probability } \ge 1 - e^{-t}$$
$$\mathbb{E}_{\widehat{P}_n}[f] \ge \mathbb{E}[f] - \sqrt{\frac{2\mathrm{Var}(f)t}{n}} - \frac{2M}{3n}t \text{ with probability } \ge 1 - e^{-t}.$$

By applying Lemma A.1 to upper bound $\mathrm{Var}(f)$ with high probability (i.e. probability at least $1 - e^{-t}$), we then find that

$$\mathbb{E}[f] \le \mathbb{E}_{\widehat{P}_n}[f] + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f)t}{n-1}} + \frac{\sqrt{2t}M^2}{n^{3/2}-n} + \frac{2+3\sqrt{2}}{3}\frac{Mt}{n} \text{ with probability } \ge 1 - 2e^{-t}. \quad (25)$$

Now, let $\{f^1, \ldots, f^N\}$ be a minimal ϵ-cover of $\mathcal{F}$ of size $N = N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})})$. Suppose that inequality (25) holds for each of the f^i. Then for any f and f^i satisfying $\left\|f - f^i\right\|_{L^\infty(\mathcal{X})} \le \epsilon$, we have

$$\begin{aligned}\mathbb{E}[f] \le \mathbb{E}[f^i] + \epsilon &\le \mathbb{E}_{\widehat{P}_n}[f^i] + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f^i)t}{n-1}} + \frac{\sqrt{2t}M^2}{n^{3/2}-n} + \frac{2+3\sqrt{2}}{3}\frac{Mt}{n} + \epsilon \\ &\le \mathbb{E}_{\widehat{P}_n}[f^i] + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(f)t}{n-1}} + \frac{\sqrt{2t}M^2}{n^{3/2}-n} + \frac{2+3\sqrt{2}}{3}\frac{Mt}{n} + \left(1 + \sqrt{\frac{2t}{n-1}}\right)\epsilon,\end{aligned}$$

where we have used that $\sqrt{\mathrm{Var}(f^i)} = \sqrt{\mathrm{Var}(f^i - f + f)} \le \sqrt{\mathrm{Var}(f^i - f)} + \sqrt{\mathrm{Var}(f)}$ for any distribution, and $\mathrm{Var}(f^i - f) \le \left\|f^i - f\right\|^2_{L^\infty(\mathcal{X})} \le \epsilon^2$. Noting that $\mathbb{E}_{\widehat{P}_n}[f^i] \le \mathbb{E}_{\widehat{P}_n}[f] + \epsilon$ gives the result.

C Efficient solutions to computing the robust expectation

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

Recall the definition of the subdifferential $\partial f(\theta) = \{g \in \mathbb{R}^d : f(\theta') \ge f(\theta) + \langle g, \theta' - \theta\rangle \text{ for all } \theta'\}$, which is simply the gradient for differentiable functions f. A standard result in convex analysis [11, Theorem VI.4.4.2] is that if the vector $p^* \in \mathbb{R}^n_+$ achieving the supremum in the definition (4) of the robust risk is unique, then

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

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

$$\nabla_\theta R_n(\theta, \mathcal{P}_n) = \sum_{i=1}^n p_i^* \nabla_\theta \ell(\theta; X_i) \text{ where } p^* = \operatorname*{argmax}_{p \in \mathcal{P}_n} \left\{\sum_{i=1}^n p_i \ell(\theta; X_i)\right\}. \quad (26)$$

In order to perform gradient descent on the risk $R_n(\theta, \mathcal{P}_n)$, then, by equation (26) we require only the computation of the worst-case distribution p^*. By taking the dual of the maximization (26), this

is an efficiently solvable convex problem; for completeness, we provide in the sequel a procedure for this computation that requires time $O(n \log n + \log \frac{1}{\epsilon} \log n)$ to compute an ϵ-accurate solution to the maximization (26). 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 [8, Chapter 9.2].

We give a detailed description of the procedure we use to compute the supremum problem (7). 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

$$\text{minimize } p^\top z \;\; \text{subject to } \frac{1}{2n} \|np - \mathbf{1}\|_2^2 \le \rho, \; p \ge 0, \; p^\top \mathbf{1} = 1.$$

We take a partial dual of this minimization problem, then maximize this dual to find the optimizing p. Introducing the dual variable $\lambda \ge 0$ for the constraint that $\frac{1}{2}\|p - \frac{1}{n}\mathbf{1}\|_2^2 \le \frac{\rho}{n}$ and performing the standard min-max swap [8] (strong duality obtains for this problem because the Slater condition is satisfied by $p = \frac{1}{n}\mathbf{1}$) yields the maximization problem

$$\underset{\lambda \ge 0}{\text{maximize}} \; f(\lambda) := \inf_p \left\{ \frac{\lambda}{2} \left\| p - \frac{1}{n}\mathbf{1} \right\|_2^2 - \frac{\lambda \rho}{n} + p^\top z \mid p \ge 0, \; \mathbf{1}^\top p = 1 \right\}. \tag{27}$$

If we can efficiently compute the infimum (27), then it is possible to binary search over λ. Recall the standard fact [11, 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 (27) 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 [9] to provide an efficient computation of the infimum (27). 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

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

and we let z^2 be the vector whose entries are z_i^2. The infimum problem (27) 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 [9], 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 [9, Figure 1] to finding the unique index i such that

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

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:

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

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 $(\eta, i) = \text{FINDSHIFT}(z, \lambda, s)$ // (Figure 4)
  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 $\lambda = \frac{1}{2}(\lambda_{\max} + \lambda_{\min})$, $(\eta, i) = \text{FINDSHIFT}(z, \lambda, s)$
SET $p_i = \left(\frac{1}{n} - \frac{1}{\lambda} z_i - \eta\right)_+$ and RETURN p

Figure 3: 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.

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 3 and 4, 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 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 4: 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 $O(\log n)$.