Spectral risk-based learning using unbounded losses

Matthew J. Holland
Osaka University

El Mehdi Haress
Université Paris-Saclay

1 INTRODUCTION

While the choice of loss function is a fundamental part of the daily workflow of most users of machine learning systems, the question of which *risk* to use receives far less attention. This is likely due to a tacit acceptance that the abstract notion of "good generalization ability" is best formulated by the expected loss $\mathbf{E}_{\mathrm{P}}\,\mathrm{L}(w;Z)$, where $Z \sim \mathrm{P}$ is a random observation, and w characterizes some decision rule. While the influential learning models of Vapnik (1999) and Haussler (1992) are centered around the expected loss, it can be argued that prioritizing *average* off-sample performance is a substantial value judgement that requires more serious consideration, both by stakeholders involved in the practical side of machine learning systems, and by the theoretician interested in providing learning algorithms with formal guarantees, stated in terms of whatever "risk" is chosen.

In the last few years, notable progress has been made in terms of learning under non-traditional risks. By far the most well-studied variant is the *conditional value-at-risk* (CVaR) of the loss distribution. Numerous applications to CVaR-based sequential learning problems have been studied (Galichet et al., 2013; Tamar et al., 2015; Prashanth et al., 2020). More recently, under convex losses, finite-sample excess (CVaR) risk bounds for stochastic gradient-based learning algorithms have been obtained under essentially any loss distribution (Holland and Haress, 2021; Soma and Yoshida, 2020), while adaptive sampling strategies have been used to improve robustness to distributional shift, without relying on convexity (Curi et al., 2020). Unfortunately, despite its practical utility, CVaR is very restrictive in terms of expressible risk preferences; all losses beyond a pre-fixed threshold are given the exact same weight. A well-known generalization is the class of *spectral risks* (Acerbi, 2002), which utilize a non-constant weighting function. This dramatically improves flexibility, but comes at the cost of more complicated form, which is expensive to estimate and difficult to optimize using traditional first-order stochastic descent methods.

To address this issue, we start by taking a derivative-free approach to learning with spectral risks. Using a stochastic smoothing technique, we first derive finite-sample excess spectral risk bounds in expectation for the proposed procedure (section 4), and show how confidence boosting can be used to obtain high-probability guarantees under loss distributions assuming just finite variance (section 5). In section 6, we propose a simple modification to the derivative-free procedure that lets us integrate gradient information from the losses for faster convergence, and we empirically verify that this procedure efficiently achieves a small spectral risk, with the interesting side-effect of out-performing empirical risk minimizers in terms of misclassification error as well, uniformly across several benchmark datasets.

2 RELATED WORK

Risk and learning While the expected value of the loss distribution is central to statistical learning theory (Haussler, 1992; Vapnik, 1999), more diverse notions of risk have been studied in broader contexts, in particular notions of financial risk (Artzner et al., 1999; Rockafellar and Uryasev, 2000; Acerbi, 2002; Ruszczyński and Shapiro, 2006) and risks which capture human psychological tendencies of aversion or affection to risk in uncertain decision-making situations (Tversky and Kahneman, 1992). In the classical theory of portfolio optimization, the mean-variance notion of risk plays a central role (Markowitz, 1952), and variance-regularized stochastic learning algorithms have been studied by Duchi and Namkoong (2019). As described earlier, originally borrowed from the financial literature, CVaR has seen many direct applications to learning problems, offers a natural interpretation to the ν-SVM algorithm (Takeda and Sugiyama, 2008), and appears less explicitly in algorithms designed to minimize worst-case losses (Shalev-Shwartz and Wexler, 2016; Fan et al., 2017). More recently, classes of generalized location-deviation risks have been studied (Lee et al., 2020; Holland, 2021), though this goes beyond the traditional setting of *coherent risks* (Artzner et al., 1999). In contrast, our study of learning under spectral risks here lets us go well beyond CVaR while still retaining the properties of coherent risks.

Derivative-free methods There are some objective functions which are differentiable, but for which computation of the derivatives is not computationally tractable. In such situations, learning algorithms which attempt to approximate first-order information using just function information are of great use; see Larson et al. (2019) for a broad survey on the topic. The basic idea that we rely upon in this work is that of designing a learning algorithm to tackle an alternative objective function which is sufficiently close to the original, but with the added benefit that (stochastic) gradient information is readily available. The 0th-order gradient estimator that we use follows the seminal work of Flaxman et al. (2004); there have been many refinements to this technique over the years (Larson et al., 2019, Sec. 4), in the form of alternative gradient estimators, but the core ideas remain the same, and the refined methods just add an additional layer of notational and expositional complexity. Plugging in more advanced estimators (Balasubramanian and Ghadimi, 2021) to our general strategy is a mechanical exercise; we use the simplest possible estimator to illustrate the efficacy of our approach in a transparent way.

Spectral risks in machine learning The research on learning with spectral risks is still very limited. Recent work from Bhat and Prashanth (2019) and Pandey et al. (2019) provides estimators for the spectral risk under sub-Gaussian and sub-Exponential loss distributions, but these results are "pointwise" in that they can only be applied to pre-fixed candidates (e.g., predictors, clusters, etc.), and do not extend to learning algorithms which consider many candidates in a data-driven fashion. Work from Khim et al. (2020) includes uniform convergence for empirical spectral risks (under the name "L-risks"), though their analysis is restricted to bounded losses, and does not lead to excess spectral risk guarantees for any particular class of learning algorithms. Our approach in this work does not build directly upon these results, since instead of a traditional empirical risk minimization (ERM) approach, we take the alternative route of optimizing a smoothed variant, whose distance from the desired risk can be readily controlled. This lets us obtain excess risk bounds for an explicit procedure (section 4, Algorithm 1), with much weaker assumptions on the underlying loss distribution.

Robustness to heavy-tailed losses A problem of importance both theoretically and in practice is that spectral risks inherit the sensitivity of CVaR to (unbounded) heavy-tailed losses (Bhat and Prashanth, 2019). This means that naive empirical estimates have extremely high variance, and the previously-cited concentration results (Pandey et al., 2019; Khim et al., 2020) no longer hold. In recent years, an active line of research has studied the problem of designing algorithms with near-optimal guarantees (in terms of the traditional risk) under heavy-tailed losses; in our section 5, we show how for an important sub-class of spectral risk tasks, we can utilize standard confidence boosting techniques (Holland and Haress, 2021), integrating them with the Algorithm 1 to obtain high-probability guarantees for a procedure that does not use first-order information, and admits heavy-tailed loss distributions.

3 PRELIMINARIES

3.1 Setup

Denoting the underlying data space by $\mathcal{Z}$, we denote by $\mathrm{L} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ a generic loss function, assumed to satisfy $\mathrm{L}(w; z) \in \mathbb{R}$ for all $w \in \mathbb{R}^d$ and $z \in \mathcal{Z}$. Our general-purpose random data is $Z \sim \mathrm{P}$, and the resulting random loss values $\mathrm{L}(w; Z)$ have a distribution function denoted by $\mathrm{F}_w(u) \coloneqq \mathrm{P}\{\mathrm{L}(w; Z) \leq u\}$, for all $w \in \mathbb{R}^d$ and $u \in \mathbb{R}$. When we make use of data-driven estimates of the distribution function, we shall denote this by $\widehat{\mathrm{F}}$. For indexing purposes, we write $[k] \coloneqq \{1, \ldots, k\}$ for any positive integer k. For any sequence (U_t) of random objects, we shall denote subsequences by $U_{[t]} \coloneqq (U_1, \ldots, U_t)$.

The traditional choice of *risk function* in loss-driven machine learning tasks is the expected value. Written explicitly, this is

$$\mathrm{R}(w) \coloneqq \mathbf{E}_{\mathrm{P}}\, \mathrm{L}(w; Z) \coloneqq \int_{\mathcal{Z}} \mathrm{L}(w; z)\, \mathrm{P}(\mathrm{d}z). \tag{1}$$

Another important risk function is the *conditional value at risk*, defined for $\beta \in [0, 1)$ by

$$\begin{aligned} \mathrm{CVaR}_\beta(w) &\coloneqq \frac{1}{1-\beta} \int_\beta^1 \mathrm{VaR}_u(w)\, \mathrm{d}u \\ &= \mathbf{E}_{\mathrm{P}}\, \mathrm{L}(w; Z)\, \mathrm{I}_{\{\mathrm{L}(w;Z) \geq \mathrm{VaR}_\beta(w)\}}, \end{aligned} \tag{2}$$

where $\mathrm{VaR}_\beta(w) \coloneqq \inf\{u : \mathrm{F}_w(u) \geq \beta\}$, the β-level quantile of $\mathrm{L}(w; Z)$. In this work, our focus will be on a class of risk functions which can be given in terms of VaR_β modulated by a user-specified density function. More concretely, let $\sigma : [0, 1] \to \mathbb{R}_+$ be a non-negative, non-decreasing function that integrates to 1. We then define the *spectral risk* of w induced by σ as

$$\mathrm{R}_\sigma(w) \coloneqq \int_0^1 \mathrm{VaR}_\beta(w)\, \sigma(\beta)\, \mathrm{d}\beta. \tag{3}$$

From the definition (2) of CVaR, we see that setting $\sigma(u) = \mathrm{I}_{\{\beta < u \leq 1\}} / (1 - \beta)$, one recovers the special case of $\mathrm{R}_\sigma(w) = \mathrm{CVaR}_\beta(w)$. A direct attack on R_σ presents difficulties, in particular with respect to computing first-order (stochastic) estimates that might in principle

drive an iterative learning algorithm. In the vein of alleviating such difficulties, using insights going back to the influential work of Flaxman et al. (2004), we introduce the *smoothed spectral risk*

$$\widetilde{\mathrm{R}}_\sigma(w) := \mathbf{E}_\nu \left[\mathrm{R}_\sigma \left(w + \gamma U \right) \right], \tag{4}$$

where $U \sim \nu$ is uniformly distributed over the unit ball $\{u \in \mathbb{R}^d : \|u\| \leq 1\}$, and the parameter controlling the degree of shift satisfies $0 < \gamma < 1$.

3.2 Basic properties

As long as the loss distribution has a positive density, spectral risks can be expressed in a more convenient form, as follows.

Lemma 1. *Let F_w be invertible and differentiable for $w \in \mathbb{R}^d$. Then, we have*

$$\mathrm{R}_\sigma(w) = \mathbf{E}_{\mathrm{P}} \, \mathrm{L}(w; Z) \, \sigma \left(\mathrm{F}_w(\mathrm{L}(w; Z)) \right), \tag{5}$$

where R_σ is the spectral risk defined in (3).

Proof. Since F_w is invertible and continuous, we have $\mathrm{VaR}_\beta(w) = \mathrm{F}_w^{-1}(\beta)$ for any $w \in \mathbb{R}^d$ and $0 < \beta < 1$. We then see that

$$\begin{aligned} &\int_0^1 \mathrm{F}_w^{-1}(\beta) \, \sigma(\beta) \, \mathrm{d}\beta \\ &\quad = \int_{-\infty}^{\infty} \mathrm{F}_w^{-1}(\mathrm{F}_w(u)) \, \sigma(\mathrm{F}_w(u)) \, \mathrm{F}_w'(u) \, \mathrm{d}u \\ &\quad = \int_{-\infty}^{\infty} u \, \sigma(\mathrm{F}_w(u)) \, \mathrm{F}_w(\mathrm{d}u), \end{aligned}$$

noting that the first equality uses integration by substitution. The right-most expression is none other than $\mathbf{E}_{\mathrm{P}} \, \mathrm{L}(w; Z) \, \sigma(\mathrm{F}_w(\mathrm{L}(w; Z)))$. □

With Lemma 1 as context, the following stochastic estimators will be of interest:

$$\mathrm{r}_\sigma(w; Z) := \mathrm{L}(w; Z) \, \sigma \left(\mathrm{F}_w(\mathrm{L}(w; Z)) \right) \tag{6}$$

$$\widehat{\mathrm{r}}_\sigma(w; Z) := \mathrm{L}(w; Z) \, \sigma \left(\widehat{\mathrm{F}}_w(\mathrm{L}(w; Z)) \right). \tag{7}$$

If the distribution function F_w were known, then access to a random sample of $Z \sim \mathrm{P}$ would immediately imply access to an unbiased estimator of R_σ. Unfortunately, in practice F_w will never be known, and thus must be estimated based on observable data. We are denoting this empirical estimator as $\widehat{\mathrm{F}}_w$. Since $\widehat{\mathrm{r}}_\sigma$ can be computed based on observable data, it will play a central role in the algorithms we study in the following section.

The following lemma gives a useful representation of any spectral risk in terms of CVaR, which using linearity of the integral lets us inherit some useful properties of the latter.

Lemma 2 (Shapiro (2013), Rmk. 3, Eqn. 42)**.** *For the spectral risk R_σ given by (3), we can write*

$$\mathrm{R}_\sigma(w) = \int_0^1 \mathrm{CVaR}_\beta(w) \, \mu_\sigma(\mathrm{d}\beta), \quad w \in \mathbb{R}^d \tag{8}$$

where μ_σ is a measure on the unit interval that does not depend on w.

Introducing the smoothed risk $\widetilde{\mathrm{R}}_\sigma$ is only going to be fruitful if it is easier to optimize than the original non-smooth risk. Fortunately, as the following result shows, it is straightforward to obtain unbiased first-order information for the smoothed risk.

Lemma 3. *In contrast with ν used in definition (4), let ν_1 denote the uniform distribution over the unit sphere $\{u \in \mathbb{R}^d : \|u\| = 1\}$, taking random direction $U \sim \nu_1$, we have*

$$\frac{d}{\gamma} \, \mathbf{E}_{\nu_1} \left[\mathrm{R}_\sigma(w + \gamma U) U \right] = \nabla \widetilde{\mathrm{R}}_\sigma(w) \tag{9}$$

for any $w \in \mathbb{R}^d$ and $0 < \gamma < 1$.

Proof. Follows from Flaxman et al. (2004, Lem. 1), using our smoothed risk (4). □

Remark 4 (Difficulties with differentiation). From the form given in Lemma 1, it is clear that using a sufficiently smooth σ, assuming we can take the derivative under the integral, then the spectral risk is indeed a differentiable function. That said, while it is differentiable in principle, we would like to emphasize to the reader that differentiation *in practice* is extremely unwieldy. Even in the ideal situation in which F_w is known, the derivative with respect to w depends on a compound of two functions that both depend on w, namely the loss $w \mapsto \mathrm{L}(w; \cdot)$ being computed, and the distribution function $w \mapsto F_w$ of the random loss $\mathrm{L}(w; Z)$. As a toy exercise that illustrates this difficulty, consider the case in which F_w represents the CDF of a Normal distribution, with mean 0 and standard deviation $\|w\|_2$. While in principle possible, many applications of the chain rule lead to complicated expressions, even in this ideal setting. Compounding this with the fact that we can never know F_w in practice, the derivative-free approach taken here provides a practical, principled, and flexible alternative.

4 GUARANTEES IN EXPECTATION ON $\mathbb{R}^d$

In this section, we specify a concrete learning algorithm, and seek excess spectral risk bounds in expectation. This procedure and its guarantees will act as a key building block to be utilized in the following section.

4.1 Algorithm analysis

Learning algorithm We essentially consider a stochastic mirror descent update, with first-order estimates using the form suggested by Lemmas 1 and 3. Making this more explicit, let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function, and let D_Φ denote the Bregman divergence induced by Φ. Modulated by positive step sizes (α_t), we generate a sequence of iterates (w_t) using the following update rule:

$$w_{t+1} = \operatorname*{arg\,min}_{w \in \mathcal{W}} \left[\langle \widehat{G}_t, w \rangle + \frac{1}{\alpha_t} \mathrm{D}_\Phi(w; w_t) \right]. \tag{10}$$

The key stochastic "gradients" used here are defined as

$$\widehat{G}_t := \frac{d}{\gamma} \widehat{\mathrm{r}}_\sigma \left(w_t + \gamma U_t; Z_t \right) U_t, \tag{11}$$

where underlying sequences (U_t) and (Z_t) are assumed to be iid, with $U_t \sim \nu_1$ and $Z_t \sim \mathrm{P}$ for all integer $t > 0$, and $\widehat{\mathrm{r}}_\sigma$ is as defined in (7). The full procedure is summarized in Algorithm 1.

Technical conditions Letting $\mathcal{W} \subset \mathbb{R}^d$ be closed, bounded, and convex, denote any minimizer of R_σ over $\mathcal{W}$ by w^*. Denote the diameter of $\mathcal{W}$, measured respectively using the underlying norm $\|\cdot\|$ and the Bregman divergence D_Φ, as $\Delta := \sup\{\|w - w'\| : w, w' \in \mathcal{W}\}$ and $\Delta_\Phi := \sup\{\mathrm{D}_\Phi(w; w') : w, w' \in \mathcal{W}\}$. Since the random perturbations may take us to points outside $\mathcal{W}$, let us define $\mathcal{C} := \{w + u : w \in \mathcal{W}, \|u\| \le 1\}$ to cover all such possibilities. Let Φ be κ-strongly convex on $\mathcal{C}$ (e.g., $\Phi(u) = \|u\|_2^2/2$, with $\kappa = 1$). On the underlying loss distribution, we assume the following moment bounds are finite:

$$\begin{aligned}
\lambda_{\mathrm{R}} &:= \sup_{v \in \mathcal{C}} \mathrm{R}(v) \\
s_1^2 &:= \sup_{v \in \mathcal{C}} \mathbf{E}_{\nu_1, \mathrm{P}} \left[\| \mathrm{r}_\sigma(v; Z) U - \mathbf{E}_{\nu_1, \mathrm{P}} \left[\mathrm{r}_\sigma(v; Z) U \right] \|^2 \right] \\
s_2^2 &:= \sup_{v \in \mathcal{C}} \mathbf{E}_{\mathrm{P}} \, |\mathrm{L}(v; Z)|^2.
\end{aligned}$$

Finally, we assume that the conditions of Lemmas 1–3 hold, the loss is such that $w \mapsto \mathrm{L}(w; Z)$ is convex and continuous on $\mathcal{C}$, and that the spectral density $\sigma(\cdot)$ is λ_σ-Lipschitz.

Theorem 5 (Spectral risk bounds in expectation)**.** *Under the preceding assumptions, let $\overline{w}_T$ be the output of Algorithm 1 run for T steps, using M points for distribution estimates, and step sizes $\alpha_t = \kappa/(\lambda_{\mathrm{R}} + 1/c_T)$ with $c_T := (\gamma/d)\sqrt{2\Delta_\Phi \kappa / (T(s_1^2 + (\lambda_\sigma s_2)^2))}$, fixed for all t. Then we have*

$$\begin{aligned}
&\mathbf{E}\left[\mathrm{R}_\sigma(\overline{w}_T) - \mathrm{R}_\sigma(w^*)\right] \\
&\quad \le 2\lambda_{\mathrm{R}}\gamma + \frac{d}{\gamma}\left[\sqrt{\frac{2\Delta_\Phi(s_1^2 + (\lambda_\sigma s_2)^2)}{T\kappa}}\right. \\
&\qquad \left. + \frac{\lambda_{\mathrm{R}}\Delta_\Phi}{T\kappa} + \lambda_\sigma \lambda_{\mathrm{R}} \Delta \sqrt{\frac{\pi}{2M}}\right]
\end{aligned}$$

for any choice of $0 < \gamma < 1$, where expectation is taken over $U_{[T]}$, $Z_{[T]}$, and the ancillary data.

The proof of Theorem 5 is composed of several simple steps, but due to its length, we just give a sketch (section 4.2), and relegate the full proof details of this and subsequent results to the supplementary materials.

Sample complexity The guarantee given by Theorem 5 is quite general, since the parameters γ, T, and M are free to be set as desired. Let us consider the important situation in which we are constrained to at most n iid samples from the data distribution P. In running Algorithm 1, for a simple and concrete example, let us set $M = \lceil \sqrt{n} \rceil$ to specify a precision level. Since each step uses $M + 1$ points, the number of steps T can thus be no greater than $n/(1 + \lceil \sqrt{n} \rceil)$, and setting $T = \lfloor n/(1 + \lceil \sqrt{n} \rceil) \rfloor$ we will always be on budget, i.e., $T(M + 1) \le n$. Plugging these values in for T and M, and subsequently minimizing the bound from Theorem 5 as a function of γ, we get $\mathbf{E}\left[\mathrm{R}_\sigma(\overline{w}_T) - \mathrm{R}_\sigma(w^*)\right] \le \varepsilon_1(n)$, where we define

$$\begin{aligned}
&\varepsilon_1(n) := \\
&2\left(2\lambda_{\mathrm{R}} d \left[\sqrt{\frac{2\Delta_\Phi(s_1^2 + (\lambda_\sigma s_2)^2)}{\lfloor n/(1 + \lceil \sqrt{n} \rceil) \rfloor \kappa}}\right.\right. \\
&\qquad \left.\left. + \frac{\lambda_{\mathrm{R}}\Delta_\Phi}{\lfloor n/(1 + \lceil \sqrt{n} \rceil) \rfloor \kappa} + \lambda_\sigma \lambda_{\mathrm{R}} \Delta \sqrt{\frac{\pi}{2\lceil \sqrt{n} \rceil}}\right]\right)^{1/2}
\end{aligned} \tag{12}$$

and thus to achieve $\mathbf{E}\left[\mathrm{R}_\sigma(\overline{w}_T) - \mathrm{R}_\sigma(w^*)\right] \le \epsilon$, the sample complexity is $\mathcal{O}(\epsilon^{-8})$.

Discussion of rates in the derivative-free literature Here we try to place the sample complexity derived from (12) into some context. For a convex objective, the main result of Flaxman et al. (2004, Thm. 1) yields a sample complexity of $\mathcal{O}(\epsilon^{-6})$ for a derivative-free update analogous to the one used here. The reason for the slower $\mathcal{O}(\epsilon^{-8})$ rate here is clear: while Flaxman et al. (2004) consider traditional risks, for our setup using spectral risks, we allocate (most) data to estimate F_{w_t} at each step, a requirement that does not arise in the traditional setting. To obtain faster rates, there are several natural routes. First, one could optimize the bound in Theorem 5 with respect to ancillary dataset size M; we set $M = \sqrt{n}$

Algorithm 1 Derivative-free stochastic mirror descent under spectral risks.

inputs: initial point $w_0 \in \mathcal{W}$, step sizes (α_t), data set size M, and max iterations T.

for $t \in \{0, \dots, T-1\}$ **do**

Get ancillary sample $\{Z'_{t,1}, \dots, Z'_{t,M}\}$, setting $\widehat{\mathrm{F}}_{w_t}(u) := (1/M) \sum_{i=1}^{M} \mathrm{I}_{\{\mathrm{L}(w_t; Z'_{t,i}) \leq u\}}$.

Sample U_t and Z_t, compute gradient $\widehat{G}_t$ via (11).

Update $w_t \mapsto w_{t+1}$ via (10).

end for

return: $\overline{w}_T := (1/T) \sum_{t=1}^{T} w_t$.

here for simplicity and readability. Second, our choice of the gradient estimator (11) was to maximize the ease of exposition; many alternative approaches have been studied over the past decade (Saha and Tewari, 2011; Belloni et al., 2015; Gasnikov et al., 2017; Balasubramanian and Ghadimi, 2021), and can readily be adapted to our problem setting to further improve the sample complexity; see Larson et al. (2019, Sec. 4.2) for a survey of relevant methods. The contribution of our work is showing a general strategy for constructing spectral-risk minimizing algorithms with guarantees (plus practical variants), which is why our results are stated using the simplest possible gradient estimator. Refining these rates further beyond our initial results is a straightforward exercise of plugging in the aforementioned estimators into our protocol.

Remark 6 (Faster rates for CVaR). Our Lipschitz assumption on σ in Theorem 5 precludes CVaR from the class of risks for which the performance guarantee holds. This is justifiable since sub-gradient information is easily computed for the special case of CVaR, and $\mathcal{O}(\epsilon^{-2})$ rates have already been proved in that restricted setting for stochastic sub-gradient algorithms (Soma and Yoshida, 2020; Holland and Haress, 2021).

4.2 Proof sketch for Theorem 5

In the detailed proof of Theorem 5, we have divided the argument into seven distinct steps covering different technical aspects of the problem. Here we provide an overview of the essential points of these steps.

Step 1. First, we establish that the modified spectral risk $\widetilde{\mathrm{R}}_\sigma$ is indeed smooth and convex. Under convex and continuous losses, both convexity and continuity are carried over by CVaR_β, which in turn is passed on to R_σ by the representation (8). This implies a Lipschitz property for R_σ, which in turn implies a Lipschitz property for the gradients of $\widetilde{\mathrm{R}}_\sigma$, i.e., smoothness. **Step 2.** Using the key link (9) between the original and modified risks, we can establish that *if* we knew the true distribution function and could thus sample $G_t := (d/\gamma)\, \mathrm{r}_\sigma(w_t + \gamma U_t; Z_t)\, U_t$ as our stochastic feedback, then this feedback is unbiased in the sense that $\mathbf{E}[G_t] = \mathbf{E}\left[\nabla \widetilde{\mathrm{R}}_\sigma(w_t)\right]$. While this ideal quantity cannot be observed, knowledge of this unbiasedness will be useful, once we have a sufficiently good approximation of the distribution function. **Step 3.** Since we have established convexity and smoothness of $\widetilde{\mathrm{R}}_\sigma$, we will eventually use this as the objective in a stochastic mirror descent program, and to set up for the analysis of those iterates, we start by bounding the differences $\widetilde{\mathrm{R}}_\sigma(w_{t+1}) - \widetilde{\mathrm{R}}_\sigma(w_t)$ in terms of $\langle \widehat{G}_t, w_{t+1} - w_t \rangle$ and the gradient "errors" $\|G_t - \widehat{G}\|$ and $\|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|$.

Controlling these three terms represents the bulk of the work done in the next two steps. Using standard mirror descent analysis techniques, combined with an argument critically utilizing the Lipschitz continuity of σ, we can control the first of these terms using the estimation error of the empirical distribution function, which enjoys sharp error bounds (**Step 4**). In **Step 5**, we note that the second of these terms only requires bounded second moments of the losses, and the third term can be controlled by taking advantage of the unbiased property established three steps earlier.

From here, a bit of cleanup is required to establish excess risk bounds in terms of the *smoothed* risk (**Step 6**). This essentially amounts to plugging in the bounds obtained in the previous two steps into the key upper bound obtained in **Step 3**, plus cleanup of telescoping sums by using basic properties of Bregman divergences; this part is rather typical for mirror descent procedures. Finally in **Step 7** we just need to derive excess spectral risk bounds from those obtained for the smoothed spectral risk, a process which is aided by the strong continuity properties of the spectral risk established in **Step 1**, plus the fact that the size of the expected norm of the noise used in smoothing is under our control. Essentially, the strong continuity properties of σ allows us to control key error terms using the error incurred by the empirical distribution function, which is congenial to control as long as we have the data for it. We make liberal use of expectations here, and obtaining high-probability control either requires stronger assumptions or a more sophisticated learning procedure. We treat this point in some detail in the next section.

5 HIGH-PROBABILITY GUARANTEES FOR UNBOUNDED LOSSES

Theorem 5 only provides guarantees in expectation, and thus it is natural to consider the output of Algorithm 1 as an inexpensive but "weak" candidate. If we split up the data, obtaining multiple weak candidates and setting aside some data for careful validation, then we can apply a robust confidence-boosting technique, as follows.

If n is our budget for sampling from P, and we want k independent candidates, run Algorithm 1 k times independently, using $\lfloor n/(k+1) \rfloor$ points each time. Denote the output of these sub-processes by $\overline{w}^{(1)}, \ldots, \overline{w}^{(k)}$. Having computed these, we still have $\lfloor n/(k+1) \rfloor$ points left, and this data will be needed to determine which of the k candidates to use. One half of this remaining data is used to construct $\widehat{\mathrm{F}}_w$, distinct from the estimates used within Algorithm 1 to get each $\overline{w}^{(j)}$. The other half, denoted Z_i'' for $i = 1, \ldots, \lfloor n/(k+1) \rfloor/2$, is used to compute a robust location estimate. As a concrete example, for each j compute

$$\widehat{\mathrm{R}}_\sigma^{(j)} := \operatorname*{arg\,min}_{a \in \mathbb{R}} \sum_{i=1}^{\lfloor n/(k+1) \rfloor/2} \rho\left(\frac{a - \mathrm{L}(\overline{w}^{(j)}; Z_i'')\, \sigma(\widehat{\mathrm{F}}_{i,j})}{b} \right), \tag{13}$$

where we have set $\widehat{\mathrm{F}}_{i,j} := \widehat{\mathrm{F}}_{\overline{w}^{(j)}}(\mathrm{L}(\overline{w}^{(j)}; Z_i''))$ for readability, ρ is a differentiable strictly convex function, and $b > 0$ is a scaling parameter. For an appropriate choice of ρ and b, this is an M-estimator of the spectral risk incurred by $\overline{w}^{(j)}$ (Catoni, 2012; Devroye et al., 2016). For each j, we introduce the key intermediate quantity

$$\overline{\mathrm{R}}_\sigma^{(j)} := \mathbf{E}_{\mathrm{P}} \left[\mathrm{L}(\overline{w}^{(j)}; Z)\, \sigma(\widehat{\mathrm{F}}_{\overline{w}^{(j)}}(\mathrm{L}(\overline{w}^{(j)}; Z))) \right]. \tag{14}$$

For comparison, let write $\mathrm{R}_\sigma^{(j)} := \mathrm{R}_\sigma(\overline{w}^{(j)})$ for the spectral risk incurred by the jth candidate. Assuming the spectral density is bounded as $\sigma(\cdot) \leq \bar{\sigma} < \infty$, then we can obtain the following upper bounds:

$$\begin{aligned}
&|\widehat{\mathrm{R}}_\sigma^{(j)} - \mathrm{R}_\sigma^{(j)}| \\
&\leq |\widehat{\mathrm{R}}_\sigma^{(j)} - \overline{\mathrm{R}}_\sigma^{(j)}| + |\overline{\mathrm{R}}_\sigma^{(j)} - \mathrm{R}_\sigma^{(j)}| \\
&\leq |\widehat{\mathrm{R}}_\sigma^{(j)} - \overline{\mathrm{R}}_\sigma^{(j)}| \\
&\qquad + \lambda_\sigma\, \mathbf{E}_{\mathrm{P}} |\mathrm{L}(\overline{w}^{(j)}; Z)| \left[\sup_{u \in \mathbb{R}} |\widehat{\mathrm{F}}_{\overline{w}^{(j)}}(u) - \mathrm{F}_{\overline{w}^{(j)}}(u)| \right] \\
&\leq \varepsilon_2(n; k, \delta) \\
&:= 2\bar{\sigma} s_2 \sqrt{\frac{2(1 + \log(2\delta^{-1}))}{\lfloor n/(k+1) \rfloor}} + \lambda_\sigma s_2 \sqrt{\frac{\log(4\delta^{-1})}{\lfloor n/(k+1) \rfloor}},
\end{aligned} \tag{15}$$

where (15) holds with probability no less than $1-\delta$, over the random draw of the data points used to compute $\widehat{\mathrm{F}}_w$ and $\widehat{\mathrm{R}}_\sigma^{(j)}$ here, conditioned on $\overline{w}^{(j)}$ (detailed proof in the appendix). Algorithmically, all we need to do is choose the best candidate based on the above robust estimates, namely

$$\overline{w}^* := \overline{w}^{(\star)}, \text{ where } \star := \operatorname*{arg\,min}_{j \in [k]} \widehat{\mathrm{R}}_\sigma^{(j)}. \tag{16}$$

This "boosted" choice enjoys a high-probability guarantee, as desired.

Theorem 7. *For confidence parameter $0 < \delta < 1$, if we set the number of weak candidates to $k = \lceil \log(2\lceil \log(\delta^{-1}) \rceil \delta^{-1}) \rceil$ and compute $\overline{w}^*$ as in (16), then we have*

$$\mathrm{R}_\sigma(\overline{w}^*) - \mathrm{R}_\sigma(w^*) \leq \mathrm{e}\, \varepsilon_1 \left(\frac{n}{k+1} \right) + 2\varepsilon_2(n; k, \delta)$$

with probability no less than $1 - 3\delta$, where ε_1 and ε_2 are as defined in (12) and (15).

Due to limited space, the proof of this result is relegated to the supplementary materials.

6 FAST IMPLEMENTATION AND EMPIRICAL ANALYSIS

The procedure outlined by Algorithm 1 yields clear formal guarantees for a wide class of spectral risks where exact gradient computations are infeasible, as described in Theorems 5 and 7. On the other hand, in the interest of practical utility, we would like to improve the slow convergence rates using even approximate first-order information, since in many cases both L and σ will be at least sub-differentiable. In this section, we outline a simple modified procedure which makes more direct use of the first-order information we have, empirically comparing it with both Algorithm 1 and traditional ERM, as natural benchmarks.

Modified procedure Issues with differentiability arise chiefly because the form of F_w is unknown. Arguably the simplest way to circumvent this difficulty is to introduce a parametric model to approximate the loss CDF. Our modified procedure takes Algorithm 1 as a starting point, and makes the following changes. First, at each step in the main loop, instead of $\widehat{\mathrm{F}}_{w_t}$, we use a folded Normal distribution, with mean and standard deviation parameters set using empirical estimates based on the ancillary sample, evaluated at w_t. Denote this parametric estimate of F_{w_t} by $\widehat{\mathrm{F}}_t$, and its derivative by $\widehat{\mathrm{f}}_t$. Next, conditioned on $\widehat{\mathrm{F}}_t$, a few applications of the chain rule lets us compute the partial

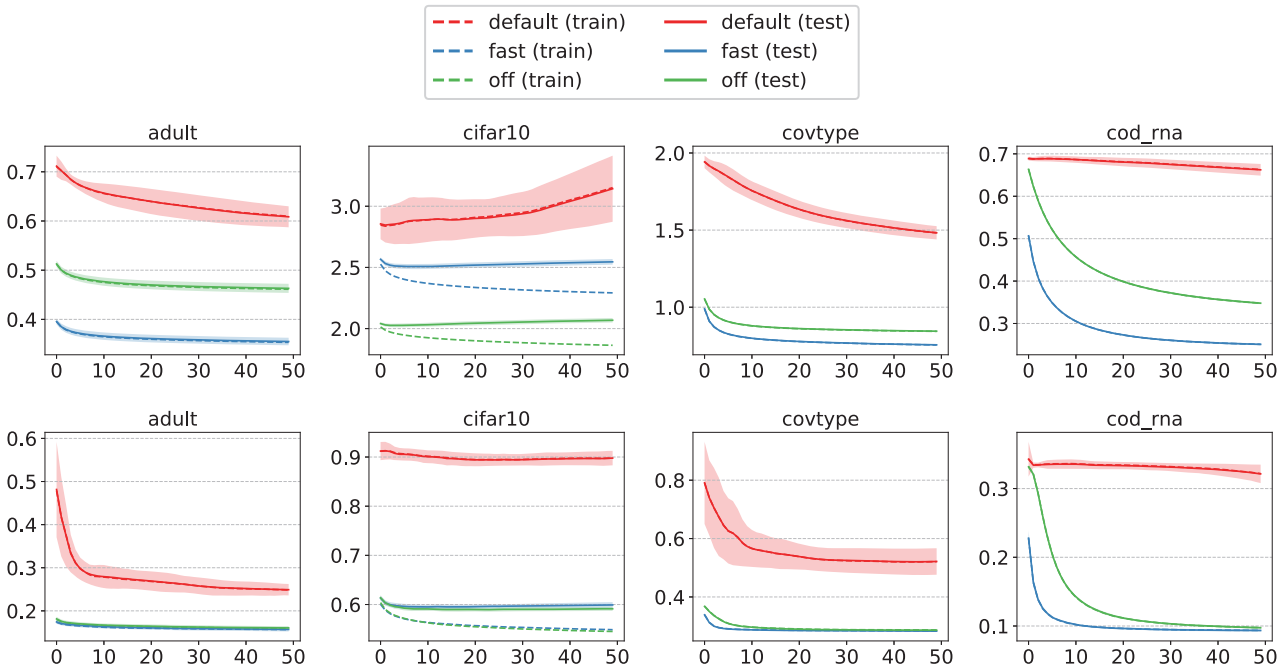

Figure 1: Error trajectories for each method and dataset, on both training (dashed) and testing (solid) data. Top row: empirical spectral risks. Bottom row: misclassification rates. Horizontal axis denotes elapsed epochs.

derivatives of $w \mapsto \mathrm{L}(w; Z)\, \sigma(\widehat{\mathrm{F}}_t(\mathrm{L}(w; Z)))$ easily. At each step t we will use the following gradient estimate:

$$\widetilde{G}_t := \Big[\sigma(\widehat{\mathrm{F}}_t(\mathrm{L}_t)) + \mathrm{L}_t\, \sigma'(\widehat{\mathrm{F}}_t(\mathrm{L}_t))\, \widehat{\mathrm{f}}_t(\mathrm{L}_t)\Big] \nabla \mathrm{L}(w_t; Z_t), \tag{17}$$

where we have written $\mathrm{L}_t := \mathrm{L}(w_t; Z_t)$ for readability. Our modified procedure is completed by using the update (10), replacing $\widehat{G}_t$ with $\widetilde{G}_t$ just specified.

Experimental design We compare three methods: derivative-free Algorithm 1 (called `default` in the figures), the modified procedure described in the previous paragraph (called `fast`), and traditional empirical risk minimization (called `off`). We mean "traditional" in terms of the risk, and thus `off` amounts to running (10) and simply replacing $\widehat{G}_t$ with the original loss gradient $\nabla \mathrm{L}(w_t; Z_t)$, and using all data for training (no ancillary set needed). All methods are run using the Euclidean norm for distance computation, and thus (10) is just a standard steepest descent update with step size α_t, plug projection onto $\mathcal{W}$. We apply each of these methods to classification tasks on a number of standard benchmark datasets, using standard multi-class logistic regression. For `default`, we fix $\alpha_t = 2\gamma/(d\sqrt{n})$, where d is the total number of parameters to be determined, and n is the number of training samples. The extra factor γ/d is to account for the coefficient in (11); this approach mirrors other derivative-free procedures (Flaxman et al., 2004, Thm. 1). For `fast` and `off`, we simply fix $\alpha_t = 2/\sqrt{n}$. These settings were selected before running any experiments. For each dataset, 10 independent trials are run, in which the full dataset is randomly shuffled before starting, with each method randomly initialized to the same point, and run for 50 epochs. Finally, as an illustrative example for our tests, we set $\sigma(\cdot)$ to the exponential risk spectrum $\sigma(u) = c\, \mathrm{e}^{-c(1-u)}/(1-\mathrm{e}^{-c})$, fixing $c = 1$, a well-established standard from the literature (Dowd and Blake, 2006; Pandey et al., 2019).

Additional details Our empirical tests have been implemented in Python (v. 3.8) with the following open-source software: matplotlib (v. 3.4.1), PyTables (v. 3.6.1), Jupyter notebook, NumPy (v. 1.20.0), and SciPy (v. 1.6.2, for special functions). See Table 1 for URLs to online documentation for each of the datasets used in our experiments. As discussed in the main text, we use multi-class logistic regression, with one linear model for each class, so the number of parameters to be determined is the number of classes (e.g., 2 for `adult`, 47 for `emnist_balanced`) multiplied by the number of input features (e.g., 105 for `adult`, 784 for `emnist_balanced`). Categorical features are given a one-hot representation, and all input features are standardized to take values on the unit interval $[0, 1]$.

Software In order to ensure our empirical analysis and results can be readily reproduced, we provide all

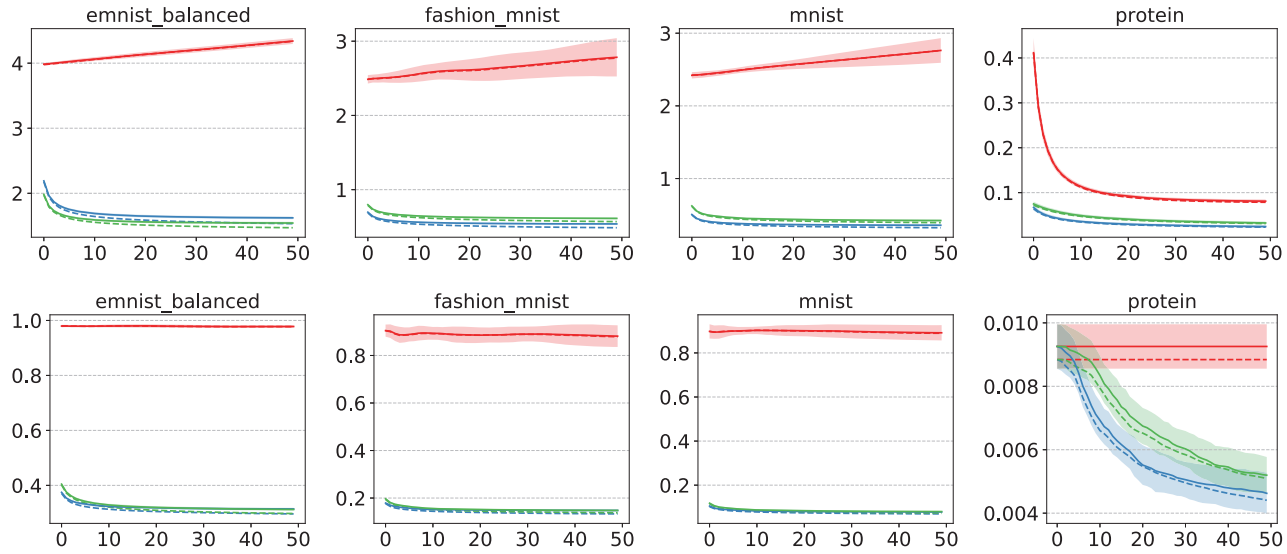

Figure 2: Completely analogous to Figure 1, for four additional datasets.

the necessary code for pre-processing the data, executing the experiments, and re-creating the figures in this paper at an following online repository.[1]

Results and discussion Plots of empirical spectral risk and misclassification rates are shown in Figures 1–2. The plotted trajectories represent averages taken over all trials, and the shaded area around the *test* error is the average ± standard deviation. The horizontal axis represents epochs, i.e., the number of passes made over the training data set. Plot titles (e.g., `cifar10`) refer to the datasets used. Additional data details are given in the appendix. One key observation that can be made is that the proposed modification `fast` achieves an appealing balance of performance in terms of spectral risk and misclassification error. Depending on the dataset, we see that without tuning the step size parameter, `off` may outperform `fast` in terms of the spectral risk, though the only stark difference appears in the case of `cifar10`, and additional testing has shown this can be mitigated with more careful step size setting. That said, it is quite remarkable that `fast` maintains a superior misclassification rate across all datasets tested. On the other hand, as suggested by the results of section 4, `default` is slow to converge, and also quite sensitive to step size settings. For simplicity and transparency we have used a fixed step size for each method, and though it should be noted that dataset-specific tuning of the step size does allow us to ensure `default` converges at close to the expected rate, the clear differences in sensitivity and speed make `fast` the first choice for practical spectral risk-based learning tasks. To further refine `default`, introducing multi-point derivative-free methods and update rules that better utilize sparse inputs (Balasubramanian and Ghadimi, 2021). As for `fast`, since the current model is quite naive with respect to the form of the loss distribution, introducing more robust modeling techniques (Lange et al., 1989) is expected to have a major impact on practical utility.

[1] https://github.com/feedbackward/spectral

7 CONCLUDING REMARKS

We have studied a derivative-free learning procedure (Algorithm 1) with excess spectral risk guarantees, under losses that may be unbounded and heavy-tailed (Theorems 5 and 7), and provided a fast implementation which on numerous real-world classification tasks has been shown to be efficient without any hyperparameter tuning. Given the existing work on spectral risk estimation (Pandey et al., 2019) and ERM for spectral risks (Khim et al., 2020), our results contribute to the literature by providing a transparent algorithmic solution for spectral risk-based learning, which is easy to implement and comes with lucid formal guarantees, plus a modified procedure that scales better to larger tasks.

Moving forward, the approach via Lemma 3 relies crucially on Stokes' theorem on $\mathbb{R}^d$, and lacks an analogue on richer spaces. Function space representations are useful in many learning methods (Dai et al., 2014; Nitanda and Suzuki, 2018), and extending Theorem 5 to general Hilbert spaces is a point of interest. How should the noise be generated? How should derivatives be defined? While a direct analogue using differential theory (e.g., Fréchet differentials (Penot, 2012)) appears diffi-

cult, key results in Malliavin calculus (Decreusefond, 2019) may open the door to a major generalization of the initial results established here.

Acknowledgements

This work was supported by the JSPS KAKENHI Grant Number 19K20342, and by JST ACT-X Grant Number JP- MJAX200O.

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Ash, R. B. and Doléans-Dade, C. A. (2000). *Probability and Measure Theory*. Academic Press, 2nd edition.

Balasubramanian, K. and Ghadimi, S. (2021). Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42.

Belloni, A., Liang, T., Narayanan, H., and Rakhlin, A. (2015). Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 240–265.

Bhat, S. P. and Prashanth, L. A. (2019). Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Optimization*, 8(3–4):231–357.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Curi, S., Levy, K. Y., Jegelka, S., and Krause, A. (2020). Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M.-F., and Song, L. (2014). Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3041–3049.

Decreusefond, L. (2019). Selected topics in Malliavin calculus. Technical report, Université Paris-Saclay. Lecture notes.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *Annals of Statistics*, 44(6):2695–2725.

Dowd, K. and Blake, D. (2006). After VaR: the theory, estimation, and insurance applications of quantile-based risk measures. *Journal of Risk and Insurance*, 73(2):193–229.

Duchi, J. and Namkoong, H. (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(1):2450–2504.

Fan, Y., Lyu, S., Ying, Y., and Hu, B. (2017). Learning with average top-k loss. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2004). Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint arXiv:cs/0408007v1*.

Galichet, N., Sebag, M., and Teytaud, O. (2013). Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *5th Asian Conference on Machine Learning (ACML 2013)*, volume 29 of *Proceedings of Machine Learning Research*, pages 245–260.

Gasnikov, A. V., Krymova, E. A., Lagunovskaya, A. A., Usmanova, I. N., and Fedorenko, F. A. (2017). Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. *Automation and Remote Control*, 78(2):224–234.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150.

Holland, M. J. (2020). Better scalability under potentially heavy-tailed feedback. *arXiv preprint arXiv:2012.07346v1*.

Holland, M. J. (2021). Learning with risks based on M-location. *arXiv preprint arXiv:2012.02424v2*.

Holland, M. J. and Haress, E. M. (2021). Learning with risk-averse feedback under potentially heavy tails. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*.

Khim, J., Leqi, L., Prasad, A., and Ravikumar, P. (2020). Uniform convergence of rank-weighted learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5254–5263.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896.

Larson, J., Menickelly, M., and Wild, S. M. (2019). Derivative-free optimization methods. *arXiv preprint arXiv:1904.11585v2*.

Lee, J., Park, S., and Shin, J. (2020). Learning bounds for risk-sensitive learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 13867–13879.

Lo, A. (2018). Demystifying the integrated tail probability expectation formula. *The American Statistician*.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.

Nitanda, A. and Suzuki, T. (2018). Functional gradient boosting based on residual network perception. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3819–3828.

Orabona, F. (2020). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213v3*.

Pandey, A. K., Prashanth, L. A., and Bhat, S. P. (2019). Estimation of spectral risk measures. *arXiv preprint arXiv:1912.10398*.

Penot, J.-P. (2012). *Calculus Without Derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer.

Prashanth, L. A., Jagannathan, K., and Kolla, R. K. (2020). Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5577–5586.

Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42.

Ruszczyński, A. and Shapiro, A. (2006). Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR W&CP*, pages 636–642.

Shalev-Shwartz, S. and Wexler, Y. (2016). Minimizing the maximal loss: How and why. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 793–801.

Shapiro, A. (2013). On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1):142–152.

Soma, T. and Yoshida, Y. (2020). Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*.

Takeda, A. and Sugiyama, M. (2008). ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1056–1063.

Tamar, A., Glassner, Y., and Mannor, S. (2015). Optimizing the CVaR via sampling. In *29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, volume 29.

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

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2nd edition.

DETAILED PROOFS

Proofs from section 4

Proof of Theorem 5. At a high level, we first control $\widetilde{\mathrm{R}}_\sigma(\overline{w}_T)$, and then using the fact that the functions $\widetilde{\mathrm{R}}_\sigma$ and R_σ are close to each other on $\mathcal{W}$, we can cast performance in terms of R_σ. This closeness depends on the perturbation factor γ; smaller is closer. On the other hand, the smoothness coefficient of $\widetilde{\mathrm{R}}_\sigma$ grows as γ gets small, leading to a natural tradeoff. We begin the proof by showing this smoothness, which enables us to control $\widetilde{\mathrm{R}}_\sigma(\overline{w}_T)$ in a straightforward manner, using an argument that relies upon well-known properties of mirror descent procedures.

Step 1: smoothness property of the smoothed spectral risk Recalling the expression (8) for R_σ, if the map $w \mapsto \mathrm{L}(w; Z)$ is convex and continuous on $\mathcal{C}$, then so is $w \mapsto \mathrm{CVaR}_\beta(w)$.[2] From (8), this immediately implies that $w \mapsto \mathrm{R}_\sigma(w)$ is convex and continuous on $\mathcal{C}$, and thus that there exists a constant $0 < \lambda_{\mathrm{R}} \leq \sup_{v \in \mathcal{C}} \mathrm{R}_\sigma(v) < \infty$ such that

$$|\,\mathrm{R}_\sigma(v) - \mathrm{R}_\sigma(v')| \leq \lambda_{\mathrm{R}} \|v - v'\| \tag{18}$$

for all $v, v' \in \mathcal{C}$.[3] Now turning our attention to the smoothed spectral risk $\widetilde{\mathrm{R}}_\sigma$, taking $U \sim \nu_1$ and any $w, w' \in \mathcal{W}$, we write the resulting noisy parameters as $W := w + \gamma U$ and $W' := w' + \gamma U$. Using the key equality (9) along with (18) just given, and the fact that $\|U\| = 1$ almost surely $[\nu_1]$, we have

$$\begin{aligned} \|\nabla \widetilde{\mathrm{R}}_\sigma(w) - \nabla \widetilde{\mathrm{R}}_\sigma(w')\| &= \frac{d}{\gamma} \left\|\mathbf{E}_{\nu_1} U \left[\mathrm{R}_\sigma(W) - \mathrm{R}_\sigma(W')\right]\right\| \\ &\leq \frac{d}{\gamma} \left|\mathrm{R}_\sigma(W) - \mathrm{R}_\sigma(W')\right| \\ &\leq \frac{d\lambda_{\mathrm{R}}}{\gamma} \|w - w'\|. \end{aligned} \tag{19}$$

As such, we can conclude that the smoothed spectral risk $\widetilde{\mathrm{R}}_\sigma$ is indeed $(d\lambda_{\mathrm{R}}/\gamma)$-smooth on $\mathcal{W}$.

Step 2: idealized stochastic gradient As an idealized counterpart to $\widehat{G}_t$, we introduce $G_t := (d/\gamma)\, \mathrm{r}_\sigma\left(w_t + \gamma U_t; Z_t\right) U_t$. This is an ideal quantity in the sense that it is the stochastic gradient that would be obtained if the true distribution function $\mathrm{F}_t := \mathrm{F}_{w_t}$ was known. Denote the ancillary datasets used in Algorithm 1 by $\boldsymbol{Z}'_t := \{Z'_{t,1}, \ldots, Z'_{t,M}\}$, for each step t, where M is the specified size. Denote sub-sequences as $U_{[t]} := (U_1, \ldots, U_t)$ for all $t > 0$ (analogously for $Z_{[t]}$ and $\boldsymbol{Z}'_{[t]}$), and write $\mathbf{E}_{[t]}$ to denote taking expectation jointly over $(U_{[t]}, Z_{[t]}, \boldsymbol{Z}'_{[t]})$. With this notation in place, note that taking expectation over all random elements, we can readily observe

$$\begin{aligned} \mathbf{E}\left[G_t\right] &= \mathbf{E}_{[t]}\, G_t \\ &= \mathbf{E}_{[t-1]}\, \mathbf{E}_{\nu_1, \mathrm{P}} \left[G_t \,|\, U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right] \\ &= \mathbf{E}_{[t-1]}\, \mathbf{E}_{\nu_1}\, \mathbf{E}_{\mathrm{P}} \left[G_t \,|\, U_{[t]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right] \\ &= \mathbf{E}_{[t-1]}\, \mathbf{E}_{\nu_1} \left(\frac{d}{\gamma}\right) \mathrm{R}_\sigma(w_t + \gamma U_t) U_t \\ &= \mathbf{E}_{[t-1]}\, \nabla \widetilde{\mathrm{R}}_\sigma(w_t) \\ &= \mathbf{E}\left[\nabla \widetilde{\mathrm{R}}_\sigma(w_t)\right]. \end{aligned} \tag{20}$$

The first and last equalities hold because G_t and w_t are independent of all random quantities with index $t+1$ or larger. The second and third equalities use the law of total expectation.[4] The rest just uses the definition of R_σ

[2] See for example Ruszczyński and Shapiro (2006, Prop. 3.1, Lem. 3.1)

[3] Since the closure of $\mathcal{C}$ is compact, continuity implies that R_σ is bounded above on $\mathcal{C}$. The Lipschitz property follows from standard results, such as Penot (2012, Prop. 3.8).

[4] See for example Ash and Doléans-Dade (2000, Thm. 5.3.3 and 5.5.4).

and the unbiased property (9). This establishes that G_t provides us with an unbiased estimate of the gradient of the smoothed spectral risk. Although the sequence (G_t) is not directly observable, this unbiasedness will be technically useful.

Step 3: setup for mirror descent analysis As an intermediate step in the overall argument, we consider stochastic minimization of $\widetilde{\mathrm{R}}_\sigma$ using the procedure specified by (10). Say we know that $\widetilde{\mathrm{R}}_\sigma$ is convex and λ-smooth on $\mathcal{C}$.[5] Taking advantage of smoothness and convexity, the following series of inequalities will make for a good starting point:

$$\begin{aligned}
&\widetilde{\mathrm{R}}_\sigma(w_{t+1}) - \widetilde{\mathrm{R}}_\sigma(w_t) \\
&\quad\le \langle \nabla \widetilde{\mathrm{R}}_\sigma(w_t), w_{t+1} - w_t\rangle + \frac{\lambda}{2}\|w_{t+1} - w_t\|^2 \\
&\quad= \langle \widehat{G}_t, w_{t+1} - w_t\rangle + \langle G_t - \widehat{G}_t, w_{t+1} - w_t\rangle + \langle \nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t, w_{t+1} - w_t\rangle + \frac{\lambda}{2}\|w_{t+1} - w_t\|^2 \\
&\quad\le \langle \widehat{G}_t, w_{t+1} - w_t\rangle + \Big(\|G_t - \widehat{G}_t\| + \|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|\Big)\|w_{t+1} - w_t\| + \frac{\lambda}{2}\|w_{t+1} - w_t\|^2 \\
&\quad\le \langle \widehat{G}_t, w_{t+1} - w_t\rangle + \frac{c}{2}\Big(\|G_t - \widehat{G}_t\| + \|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|\Big)^2 + \left(\frac{\lambda}{2} + \frac{1}{2c}\right)\|w_{t+1} - w_t\|^2 \\
&\quad\le \langle \widehat{G}_t, w_{t+1} - w_t\rangle + c\Big(\|G_t - \widehat{G}_t\|^2 + \|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|^2\Big) + \left(\lambda + \frac{1}{c}\right)\frac{\mathrm{D}_\Phi(w_{t+1}; w_t)}{\kappa}.
\end{aligned} \tag{21}$$

The first inequality uses a basic property of functions with Lipschitz-continuous gradients.[6] The second inequality is just Cauchy-Schwarz. The third inequality uses the elementary fact $2ab \le ca^2 + b^2/c$ for any $c > 0$. The final inequality makes use of the fact that $(a+b)^2 \le 2(a^2+b^2)$, and the fact that κ-strong convexity of Φ implies $\mathrm{D}_\Phi(u; v) \ge (\kappa/2)\|u - v\|^2$.

Step 4: bounding intermediate terms Taking the first term in (21), fixing any $\widetilde{w}^* \in \mathbb{R}^d$ we trivially have

$$\langle \widehat{G}_t, w_{t+1} - w_t\rangle = \langle \widehat{G}_t, w_{t+1} - \widetilde{w}^*\rangle + \langle G_t, \widetilde{w}^* - w_t\rangle + \langle \widehat{G}_t - G_t, \widetilde{w}^* - w_t\rangle. \tag{22}$$

Taking the right-hand side one term at a time, the first term is bounded by

$$\langle \widehat{G}_t, w_{t+1} - \widetilde{w}^*\rangle \le \frac{A_t}{\alpha_t} \coloneqq \frac{\mathrm{D}_\Phi(\widetilde{w}^*; w_t) - \mathrm{D}_\Phi(w_{t+1}; w_t) - \mathrm{D}_\Phi(\widetilde{w}^*; w_{t+1})}{\alpha_t}, \tag{23}$$

a fact which holds from standard mirror descent analysis.[7] Next, taking expectation over the second term, using (20) and the convexity of $\widetilde{\mathrm{R}}_\sigma$, we have

$$\mathbf{E}\left[\langle G_t, \widetilde{w}^* - w_t\rangle + \widetilde{\mathrm{R}}_\sigma(w_t)\right] = \mathbf{E}\left[\langle \nabla \widetilde{\mathrm{R}}_\sigma(w_t), \widetilde{w}^* - w_t\rangle + \widetilde{\mathrm{R}}_\sigma(w_t)\right] \le \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*). \tag{24}$$

Finally, to deal with the remaining gradient difference term, note that

$$\begin{aligned}
\|\widehat{G}_t - G_t\| &\le \frac{d}{\gamma}\|U_t\|\,\mathrm{L}_t\,|\,\sigma(\mathrm{F}_t(\mathrm{L}_t)) - \sigma(\widehat{\mathrm{F}}_t(\mathrm{L}_t))| \\
&\le \left(\frac{d\lambda_\sigma\,\mathrm{L}_t}{\gamma}\right)\sup_{u\in\mathbb{R}}|\,\mathrm{F}_t(u) - \widehat{\mathrm{F}}_t(u)|.
\end{aligned}$$

Write $\mathbf{E}'_t$ to denote taking expectation with respect to $\boldsymbol{Z}'_t$, and for readability, write the distribution function estimation error as $\|\,\mathrm{F}_t - \widehat{\mathrm{F}}_t\,\| \coloneqq \sup_{u\in\mathbb{R}}|\,\mathrm{F}_t(u) - \widehat{\mathrm{F}}_t(u)|$. If we take the expectation of the inequality just derived

[5] In (19) we have already been proved this holds with $\lambda = d\lambda_{\mathrm{R}}/\gamma$.
[6] See for example Nesterov (2004, Thm. 2.1.5).
[7] See Bubeck (2015, Ch. 4, 6) or Orabona (2020, Ch. 6) for a highly readable background.

and use Cauchy-Schwarz, we obtain

$$\begin{aligned}
\mathbf{E}\left[\langle \widehat{G}_t - G_t, \widetilde{w}^* - w_t\rangle\right] &\leq \mathbf{E}\left[\left(\frac{d\lambda_\sigma \, \mathrm{L}_t}{\gamma}\right) \|\widetilde{w}^* - w_t\| \| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \|\right] \\
&= \mathbf{E}_{[t]}\left[\left(\frac{d\lambda_\sigma \, \mathrm{L}_t}{\gamma}\right) \|\widetilde{w}^* - w_t\| \| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \|\right] \\
&= \left(\frac{d\lambda_\sigma}{\gamma}\right) \mathbf{E}\left[\mathbf{E}'_t\left[\mathrm{L}_t \, \|\widetilde{w}^* - w_t\| \| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \| \;\middle|\; U_{[t-1]}, Z_{[t]}, \boldsymbol{Z}'_{[t-1]}\right]\right] \\
&= \left(\frac{d\lambda_\sigma}{\gamma}\right) \mathbf{E}\left[\mathrm{L}_t \, \|\widetilde{w}^* - w_t\| \, \mathbf{E}'_t\left[\| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \| \;\middle|\; U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right]\right].
\end{aligned}$$

The above equalities follow from applying the law of total expectation and noting that conditioned on $U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}$, w_t is no longer random, and conditioned on $U_{[t-1]}, Z_{[t]}, \boldsymbol{Z}'_{[t-1]}$, L_t is no longer random. To clean up this upper bound, first note that

$$\begin{aligned}
\mathbf{E}'_t\left[\| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \| \;\middle|\; U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right] &= \int_0^\infty \mathbf{P}\left\{\| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \| > \varepsilon \;\middle|\; U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right\} \mathrm{d}\varepsilon \\
&\leq 2\int_0^\infty \exp\left(-2M\varepsilon^2\right) \mathrm{d}\varepsilon \\
&= \sqrt{\frac{\pi}{2M}}.
\end{aligned}$$

The first equality is a basic probability result.[8] The inequality is just an application of the refined DKW inequality.[9] In a similar fashion, using Δ to bound the diameter of the hypothesis class $\mathcal{W}$, we have that

$$\begin{aligned}
\mathbf{E}\left[\langle \widehat{G}_t - G_t, \widetilde{w}^* - w_t\rangle\right] &\leq \left(\frac{d\lambda_\sigma}{\gamma}\right)\sqrt{\frac{\pi}{2n}} \, \mathbf{E} \, \mathrm{L}_t \, \|\widetilde{w}^* - w_t\| \\
&\leq \left(\frac{d\lambda_\sigma \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2n}} \, \mathbf{E}\left[\mathbf{E}_{\mathrm{P}}\left[\mathrm{L}(w_t; Z) \,|\, U_{[t-1]}, Z_{[t-1]}, \boldsymbol{Z}'_{[t-1]}\right]\right] \\
&= \left(\frac{d\lambda_\sigma \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2n}} \, \mathbf{E}\left[\mathrm{R}(w_t)\right] \\
&\leq \left(\frac{d\lambda_\sigma \lambda_{\mathrm{R}} \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2n}}. \qquad (25)
\end{aligned}$$

The final inequality uses the definition of λ_{R} and the fact that $\mathcal{W} \subset \mathcal{C}$. This covers the first term in (21).

Step 5: more intermediate terms For the second term in (21), we need control of $\mathbf{E}\,\|G_t - \widehat{G}_t\|^2$ and $\mathbf{E}\,\|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|^2$. As a simple bound on the first of these, noting that $\| \, \mathrm{F}_t - \widehat{\mathrm{F}}_t \, \| \leq 1$, we have

$$\mathbf{E}\,\|G_t - \widehat{G}_t\|^2 \leq \left(\frac{d\lambda_\sigma}{\gamma}\right)^2 \sup_{v \in \mathcal{C}} \mathbf{E}_{\mathrm{P}} \,|\, \mathrm{L}(v; Z)|^2.$$

For the remaining term, we have

$$\begin{aligned}
\mathbf{E}_{\nu_1,\mathrm{P}} \, \|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|^2 &= \left(\frac{d}{\gamma}\right)^2 \mathbf{E}_{\nu_1,\mathrm{P}}\left[\|\mathbf{E}_{\nu_1,\mathrm{P}}\left[\mathrm{r}_\sigma(w_t + \gamma U; Z)U\right] - \mathrm{r}_\sigma(w_t + \gamma U; Z)U\|^2\right] \\
&\leq \left(\frac{d}{\gamma}\right)^2 \sup_{v \in \mathcal{C}} \mathbf{E}_{\nu_1,\mathrm{P}}\left[\|\mathbf{E}_{\nu_1,\mathrm{P}}\left[\mathrm{r}_\sigma(v; Z)U\right] - \mathrm{r}_\sigma(v; Z)U\|^2\right].
\end{aligned}$$

The preceding inequality holds because $0 < \gamma < 1$ implies $w_t + \gamma U \in \mathcal{C}$ almost surely $[\nu_1]$. Taking expectation over all elements and using the definitions of s_1 and s_2, we have

$$\mathbf{E}\left[c\left(\|G_t - \widehat{G}_t\|^2 + \|\nabla \widetilde{\mathrm{R}}_\sigma(w_t) - G_t\|^2\right)\right] \leq c\left(\frac{d}{\gamma}\right)^2 \left((\lambda_\sigma s_2)^2 + s_1^2\right). \qquad (26)$$

[8] See Lo (2018) for a lucid elementary background on this fact.
[9] See for example Kosorok (2008, Thm. 11.6).

Step 6: cleanup to bound smoothed spectral risk To start the cleanup process, taking inequalities (23)–(26) back to (21) and taking expectation, we can immediately deduce

$$\mathbf{E}\left[\widetilde{\mathrm{R}}_\sigma(w_{t+1}) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*)\right] \leq \mathbf{E}\left[\frac{A_t}{\alpha_t} + \left(\lambda + \frac{1}{c}\right)\frac{\mathrm{D}_\Phi(w_{t+1}; w_t)}{\kappa}\right] + \left(\frac{d\lambda_\sigma \lambda_{\mathrm{R}} \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2M}} + c\left(\frac{d}{\gamma}\right)^2 \left(s_1^2 + (\lambda_\sigma s_2)^2\right).$$

For the first term in the preceding inequality, since A_t is composed of a difference of Bregman divergences, note that

$$\begin{aligned}
&\frac{A_t}{\alpha_t} + \left(\lambda + \frac{1}{c}\right)\frac{\mathrm{D}_\Phi(w_{t+1}; w_t)}{\kappa} \\
&\qquad = \frac{\mathrm{D}_\Phi(\widetilde{w}^*; w_t) - \mathrm{D}_\Phi(\widetilde{w}^*; w_{t+1})}{\alpha_t} + \mathrm{D}_\Phi(w_{t+1}; w_t)\left(\frac{1}{\kappa}\left(\frac{1}{c} + \lambda\right) - \frac{1}{\alpha_t}\right) \\
&\qquad = \frac{\mathrm{D}_\Phi(\widetilde{w}^*; w_t) - \mathrm{D}_\Phi(\widetilde{w}^*; w_{t+1})}{\alpha(c)}.
\end{aligned} \tag{27}$$

The last equality holds via the setting of $\alpha_t = \alpha(c) := \kappa(\lambda + 1/c)^{-1}$ for all t, causing the extra term to vanish. Next, leveraging Jensen's inequality and cancelling terms via the telescoping sum, we have

$$\begin{aligned}
&\mathbf{E}\left[\widetilde{\mathrm{R}}_\sigma\left(\frac{1}{T}\sum_{t=1}^T w_t\right) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*)\right] \\
&\leq \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^T \left(\widetilde{\mathrm{R}}_\sigma(w_t) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*)\right)\right] \\
&\leq \frac{\mathrm{D}_\Phi(\widetilde{w}^*; w_1) - \mathrm{D}_\Phi(\widetilde{w}^*; w_{T+1})}{T\alpha(c)} + \left(\frac{d\lambda_\sigma \lambda_{\mathrm{R}} \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2M}} + c\left(s_1^2 + (\lambda_\sigma s_2)^2\right) \\
&\leq \frac{\Delta_\Phi}{T\kappa}\left(\lambda + \frac{1}{c}\right) + \left(\frac{d\lambda_\sigma \lambda_{\mathrm{R}} \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2M}} + c\left(\frac{d}{\gamma}\right)^2 \left(s_1^2 + (\lambda_\sigma s_2)^2\right).
\end{aligned}$$

Minimizing the preceding upper bound with respect to $c > 0$, one sets

$$c = \left(\frac{\gamma}{d}\right)\sqrt{\frac{2\Delta_\Phi \kappa}{T(s_1^2 + (\lambda_\sigma s_2)^2)}}$$

and obtains the bound

$$\mathbf{E}\left[\widetilde{\mathrm{R}}_\sigma\left(\frac{1}{T}\sum_{t=1}^T w_t\right) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*)\right] \leq \left(\frac{d}{\gamma}\right)\sqrt{\frac{2\Delta_\Phi(s_1^2 + (\lambda_\sigma s_2)^2)}{T\kappa}} + \frac{\lambda\Delta_\Phi}{T\kappa} + \left(\frac{d\lambda_\sigma \lambda_{\mathrm{R}} \Delta}{\gamma}\right)\sqrt{\frac{\pi}{2M}}. \tag{28}$$

Again, we remark that this holds for any fixed choice of $\widetilde{w}^*$.

Step 7: guarantees in terms of spectral risk Using (28) we have a bound in expectation on the smoothed spectral risk $\widetilde{\mathrm{R}}_\sigma$ incurred by the (averaged) learning algorithm (10), so it remains for us to relate this to the original objective of interest, namely the spectral risk R_σ. Denote a minimizer of this objective by $w^* \in \arg\min_{w\in\mathcal{W}} \mathrm{R}_\sigma(w)$, and now let us fix $\widetilde{w}^*$ that appears in (28) to be optimal in terms of $\widetilde{\mathrm{R}}_\sigma$, that is, let $\widetilde{w}^* \in \arg\min_{w\in\mathcal{W}} \widetilde{\mathrm{R}}_\sigma$ hold. Using this optimality and continuity properties of convex R_σ, we see that

$$\begin{aligned}
\mathrm{R}_\sigma(\overline{w}_T) - \mathrm{R}_\sigma(w^*) &= \left[\mathrm{R}_\sigma(\overline{w}_T) - \widetilde{\mathrm{R}}_\sigma(\overline{w}_T)\right] + \left[\widetilde{\mathrm{R}}_\sigma(\overline{w}_T) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*)\right] + \left[\widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*) - \mathrm{R}_\sigma(w^*)\right] \\
&\leq 2\sup_{w\in\mathcal{W}}\left|\mathrm{R}_\sigma(w) - \widetilde{\mathrm{R}}_\sigma(w)\right| + \widetilde{\mathrm{R}}_\sigma(\overline{w}_T) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*) \\
&= 2\sup_{w\in\mathcal{W}}\left|\mathbf{E}_\nu(\mathrm{R}_\sigma(w) - \mathrm{R}_\sigma(w + \gamma U))\right| + \widetilde{\mathrm{R}}_\sigma(\overline{w}_T) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*) \\
&\leq 2\lambda_{\mathrm{R}}\gamma + \widetilde{\mathrm{R}}_\sigma(\overline{w}_T) - \widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*).
\end{aligned} \tag{29}$$

The first inequality follows due to the optimality of $\widetilde{w}^*$, which implies $\widetilde{\mathrm{R}}_\sigma(\widetilde{w}^*) \leq \widetilde{\mathrm{R}}_\sigma(w^*)$. The second equality follows from the definition of $\widetilde{\mathrm{R}}_\sigma$. The last inequality follows from (18) and the fact that $\mathbf{E}_\nu \|U\| \leq 1$. Taking expectation of (29), a direct application of the bound (28) with λ set according to (19) yields the desired result. □

Proofs from section 5

Proof of Theorem 7. We start by proving inequality (15), namely the key validation error bound. After bounding $|\widehat{\mathrm{R}}_\sigma^{(j)} - \mathrm{R}_\sigma^{(j)}|$ by the two difference terms, the second inequality follows immediately from the definition of the spectral risk and the intermediate quantity $\overline{\mathrm{R}}_\sigma^{(j)}$, using the λ_σ-Lipschitz property of σ to get the error in terms of the error between distribution functions.

The next step (leading to (15)) is comprised of a few parts. First, using Hölder's inequality, for any $w \in \mathcal{C}$ we have $\mathbf{E}_\mathrm{P}|\,\mathrm{L}(w;Z)| \leq \sqrt{\mathbf{E}_\mathrm{P}|\,\mathrm{L}(w;Z)|^2} \leq s_2$, by definition of s_2. Next, for any fixed w, the DKW inequality (Kosorok, 2008, Thm. 11.6) implies

$$\mathbf{P}\left\{\sup_u |\,\widehat{\mathrm{F}}_w(u) - \mathrm{F}_w(u)| > \varepsilon\right\} \leq 2\exp(-2\varepsilon^2\lfloor n/(k+1)\rfloor).$$

Thus, conditioned on $\overline{w}^{(j)}$, the bound on the second term in (15) holds with probability no less than $1-\delta/2$, over the random draw of the points used to compute the estimate $\widehat{\mathrm{F}}_w$. This is the first "good event" of interest.

The second good event is with respect to the remaining data $\{Z_i''\}$ used to compute the spectral risk estimates. Let us denote the variance of the weighted loss by

$$v_\sigma^{(j)} := \mathrm{var}_\mathrm{P}\left[\mathrm{L}(\overline{w}^{(j)};Z)\,\sigma(\widehat{\mathrm{F}}_{\overline{w}^{(j)}}(\mathrm{L}(\overline{w}^{(j)};Z)))\right].$$

Conditioning on $\widehat{\mathrm{F}}_w$ and $\overline{w}^{(j)}$ for the moment, standard concentration inequalities for M-estimators tell us that

$$|\widehat{\mathrm{R}}_\sigma^{(j)} - \overline{\mathrm{R}}_\sigma^{(j)}| \leq 2\sqrt{\frac{2v_\sigma^{(j)}(1+\log(2\delta^{-1}))}{\lfloor n/(k+1)\rfloor}} \tag{30}$$

holds with probability no less than $1-\delta/2$; see for example Catoni (2012) or Devroye et al. (2016) for typical examples of ρ and b settings. To get a bound free of the elements being conditioned upon, note that the variance of the weighted loss can be bounded as

$$\mathrm{var}_\mathrm{P}\,\mathrm{L}(w;Z)\,\sigma(\widehat{\mathrm{F}}_w(\mathrm{L}(w;Z))) \leq \bar{\sigma}^2\,\mathbf{E}_\mathrm{P}|\,\mathrm{L}(w;Z)|^2 \leq \bar{\sigma}^2 s_2^2 < \infty.$$

We can thus bound $v_\sigma^{(j)} \leq \bar{\sigma}^2 s_2^2$ in (30), and this is our second good event of interest. Taking a union bound of these two "good events" (each with probability at least $1-\delta/2$), we obtain (15) with probability at least $1-\delta$, as desired.

With inequality (15) in hand for each of the sub-processes indexed by $j = 1,\ldots,k$, we can combine this with the key learning guarantees in expectation provided by Theorem 5. In particular, we use the excess expected spectral risk bound $\varepsilon_1(\cdot)$ in (12), but this time passed a sample of size $n/(k+1)$, since that is all that each sub-process (each independent run of Algorithm 1) is allocated. The desired result then follows quite mechanically using a generic robust confidence boosting argument, as follows. First, we plug in (12) and (15) to (Holland and Haress, 2021, Lem. 9) to obtain the desired good event for general k that holds with probability no less than $1-k\delta-\mathrm{e}^{-k}$. To clean up this probability just requires careful setting of the number of partitions; defining $k_\delta := \lceil\log(\delta^{-1})\rceil$ and $\delta_* := \delta/(2k_\delta)$ for any $0<\delta<1$, in the theorem statement we set $k = k_{\delta^*} = \lceil\log(2\lceil\log(\delta^{-1})\rceil\delta^{-1})\rceil$, under which a straightforward but tedious argument shows that such as a setting of k implies[10]

$$1-k\delta-\mathrm{e}^{-k} \geq 1-3\delta.$$

This high-probability event using $k = k_{\delta_*}$ is precisely the result in our theorem statement. □

DATASET INFORMATION

In Table 1, we have included names and URLs of the datasets used in our empirical tests.

[10] See Holland (2020, Proof of Thm. 7) for all the details.

Dataset	URL
adult	https://archive.ics.uci.edu/ml/datasets/Adult
cifar10	https://www.cs.toronto.edu/~kriz/cifar.html
cod_rna	https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html
covtype	https://archive.ics.uci.edu/ml/datasets/covertype
emnist_balanced	https://www.nist.gov/itl/products-and-services/emnist-dataset
fashion_mnist	https://github.com/zalandoresearch/fashion-mnist
mnist	http://yann.lecun.com/exdb/mnist/
protein	https://www.kdd.org/kdd-cup/view/kdd-cup-2004/Data

Table 1: Benchmark dataset summary.