
Robust variance-regularized risk minimization with concomitant scaling

Matthew J. Holland
Osaka University

Abstract

Under losses which are potentially heavy-tailed, we consider the task of minimizing sums of the loss mean and standard deviation, without trying to accurately estimate the variance. By modifying a technique for variance-free robust mean estimation to fit our problem setting, we derive a simple learning procedure which can be easily combined with standard gradient-based solvers to be used in traditional machine learning workflows. Empirically, we verify that our proposed approach, despite its simplicity, performs as well or better than even the best-performing candidates derived from alternative criteria such as CVaR or DRO risks on a variety of datasets.

1 INTRODUCTION

Traditionally, the “textbook definition” of a statistical machine learning problem is formulated in terms of making decisions which minimize the expected value of a random loss (Devroye et al., 1996; Mohri et al., 2012; Vapnik, 1999). More precisely, the traditional setup has us minimize $\mathbf{E}_\mu \mathbf{L}(h)$ with respect to a decision h , where we denote random losses as $\mathbf{L}(h) := \ell(h; \mathbf{Z})$, with a random data point $\mathbf{Z} \sim \mu$, and $\ell(\cdot)$ is a loss function assigning real values to (decision, data) pairs. This problem class is very general in that it covers a wide range of learning problems both supervised and unsupervised, but it is limited in the sense that it only aspires to be optimal *on average*, with no guarantees for other aspects of performance such as loss deviations, resilience to worst-case examples and distribution shift, sub-population disparity, and class-balanced error. While it is sometimes possible to account for these issues by modifying the base loss function ℓ (e.g.,

logit-adjusted softmax cross-entropy for balanced error (Menon et al., 2021)), there is a growing literature looking at principled, systematic modifications to the “risk,” i.e., a non-random numerical property of the *distribution* of $\mathbf{L}(h)$ to be optimized in h , leaving the base loss $\ell(\cdot)$ fixed. Some prominent examples are weighted sums of loss quantiles (Medina and Yang, 2021), distributionally robust optimization (DRO) risk (Hashimoto et al., 2018), conditional value-at-risk (CVaR) (Curi et al., 2020), tilted risk (Li et al., 2021), and more general optimized certainty equivalent (OCE) risks (Lee et al., 2020), among others. It is well-known that many risks can be expressed in terms of location-deviation sums, with the canonical example being a weighted sum of the loss mean and standard deviation (or variance) (Rockafellar and Uryasev, 2013, §2). We refer the reader to some recent surveys (Holland and Tanabe, 2023; Hu et al., 2022; Royset, 2022) for more general background on developments in learning criteria.

In this work, the criterion of interest is the mean loss regularized by standard deviation (SD), when losses are allowed to be heavy-tailed. More formally, we allow for heavy tails in the sense that all we assume is that the second moment $\mathbf{E}_\mu |\mathbf{L}(h)|^2$ is finite, and the ultimate objective of interest is the *mean-SD* criterion

$$\text{MS}_\mu(h; \lambda) := \mathbf{E}_\mu \mathbf{L}(h) + \sqrt{\lambda \mathbf{V}_\mu \mathbf{L}(h)} \quad (1)$$

with loss variance denoted by $\mathbf{V}_\mu \mathbf{L} := \mathbf{E}_\mu (\mathbf{L} - \mathbf{E}_\mu \mathbf{L})^2$, and weighting parameter $\lambda \geq 0$. This mean-SD objective (1) and its mean-variance counterpart have a long history in the literature on decision making under uncertainty, including the influential work of Markowitz (1952) on optimal portfolio selection. In the context of machine learning, it is well-known that one can obtain “fast rate” bounds on the expected loss when variance is small (see (Duchi and Namkoong, 2019, §1)), though the problem of actually ensuring that loss deviations are sufficiently small is an entirely separate matter. In this direction, Maurer and Pontil (2009) bound the (population) expected loss using a weighted sum of the *sample* mean and standard deviation. Their “sample variance penalized” objective is convenient to compute and can be used to guarantee fast rates in theory, but a

lack of convexity makes it hard to minimize in practice. A convex approximation is developed by [Duchi and Namkoong \(2019\)](#), who show that a sub-class of (empirical) DRO risks can be used to approximate the sample mean-SD objective, again yielding fast rates when the (population) variance is small enough. The critical limitation to this approach is poor guarantees under heavy-tailed losses; while we gain in terms of convexity, the empirical DRO risk of ([Duchi and Namkoong, 2019](#)) is at least as sensitive to outliers as the naive empirical objective (i.e., directly minimizing the sample mean and SD), which is already known to result in highly sub-optimal performance guarantees under heavy tails ([Brownlees et al., 2015](#); [Devroye et al., 2016](#); [Hsu and Sabato, 2016](#)). Recent work by [Zhai et al. \(2021\)](#) studies a natural strategy for robustifying the DRO objective (called DORO), which discards a specified fraction of the largest losses. While the impact of outliers can be reduced under the right setting of DORO, their approach is limited to non-negative losses, and the impact that such one-sided trimming has on the resulting mean-SD sum, our ultimate object of interest, is unknown.

With this context in mind, in this paper we propose a new approach to robustly minimize the objective (1) under heavy-tailed losses, without *a priori* knowledge of anything but the fact that variance is finite. Our key technique is based on extending a convex program of [Sun \(2021\)](#) from one-dimensional mean estimation to our mean-variance objective $\text{MS}_\mu(h; \lambda)$ under general losses. After some motivating background points in §2.1–§2.2, we describe our basic approach and summarize our contributions in §2.3–§2.4. Theoretical analysis comes in §3, and based upon formal properties of the proposed objective function, we derive a general-purpose procedure summarized in Algorithm 1, and tested empirically in §4. Our main finding is that the simple algorithm we derive works remarkably well on both simulated and real-world datasets without any fine-tuning, despite sacrificing the convexity enjoyed by procedures based on criteria such as CVaR and DRO. All detailed proofs are relegated to the appendix. Software and notebooks to reproduce all results in this paper are provided in an online repository.¹

2 BACKGROUND

Before we describe our proposed approach to the mean-SD task described in §1, we start with a much simpler problem, namely the task of robust mean estimation. This will allow us to highlight key technical points from the literature which provide both conceptual and technical context for our proposal. Key points from

the existing literature are introduced in §2.1–§2.2, and building upon this we introduce our method in §2.3–§2.4.

2.1 Robust mean estimation

Let X be a random variable. For the moment, our goal will be to construct an accurate empirical estimate of the mean $\mathbf{E}_\mu X$, assuming only that the variance $\mathbf{V}_\mu X = \mathbf{E}_\mu X^2 - (\mathbf{E}_\mu X)^2$ is both defined and finite. We assume access to an independent and identically distributed (IID) sample X_1, \dots, X_n . Since higher-order moments may be infinite, the tails of X may be “heavy” and decidedly non-Gaussian, causing problems for the usual empirical mean. This problem setting is now very well-understood; see [Lugosi and Mendelson \(2019\)](#) for an authoritative reference. One very well-known approach is to use M-estimators ([Huber and Ronchetti, 2009](#)), namely to design an estimator $A_n \approx \mathbf{E}_\mu X$ satisfying

$$A_n \in \arg \min_{a \in \mathbb{R}} \frac{b}{n} \sum_{i=1}^n \rho \left(\frac{X_i - a}{b} \right) \quad (2)$$

where $\rho: \mathbb{R} \rightarrow \mathbb{R}_+$ is a function that is approximately quadratic near zero, but grows more slowly in the limit, i.e., large deviations are penalized in a sub-quadratic manner, where “large” is relative to the scaling parameter $b > 0$, used to control bias. When $\rho(\cdot)$ is convex, differentiable, and the solution set is non-empty, the condition (2) is equivalent to

$$\frac{1}{n} \sum_{i=1}^n \rho' \left(\frac{X_i - A_n}{b} \right) = 0 \quad (3)$$

and when the derivative $\rho'(\cdot)$ is bounded on \mathbb{R} such that

$$-\log(1 - x + \gamma x^2) \leq \rho'(x) \leq \log(1 + x + \gamma x^2) \quad (4)$$

for some constant $0 < \gamma < \infty$, then the approach of [Catoni \(2012\)](#) tells us that when b^2 scales with $\mathbf{V}_\mu X/n$, the deviations $|A_n - \mathbf{E}_\mu X|$ enjoy sub-Gaussian tails, namely upper bounds of the order $\mathcal{O}(\sqrt{\log(1/\delta)} \mathbf{V}_\mu X/n)$ with probability at least $1 - \delta$. Under these weak assumptions, such guarantees are essentially optimal ([Devroye et al., 2016](#)). While an important result, in practice the need for knowledge of $\mathbf{V}_\mu X$ is a significant limitation, since without finite higher-order moments, it is not plausible to obtain variance estimates with such sub-Gaussian guarantees. There do exist other robust estimators such as median-of-means ([Lugosi and Mendelson, 2019, §2.1](#)) which do not require variance information, and this illustrates the fact that knowledge of the variance is sufficient, although *not* necessary, for sub-Gaussian mean estimation under heavy tails.

¹<https://github.com/feedbackward/bdd-mv>

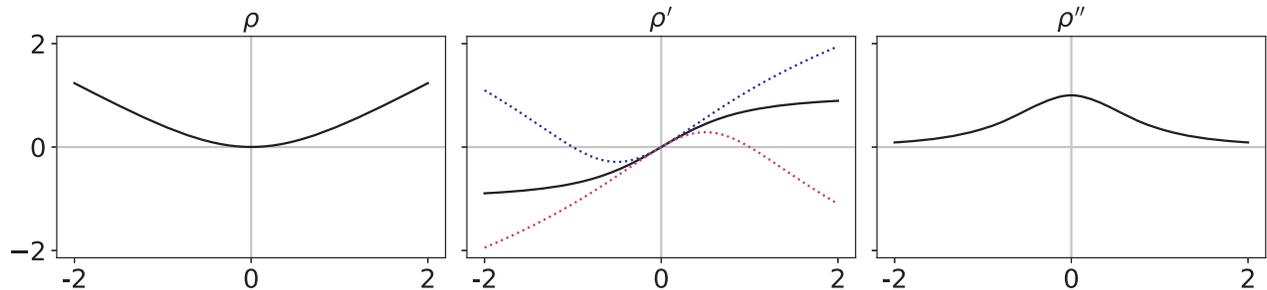


Figure 1: From left to right, we plot the graphs of $\rho(\cdot)$, $\rho'(\cdot)$, and $\rho''(\cdot)$ with ρ as in (6). In the middle plot, the dotted curves represent the upper (blue) and lower (dark pink) bounds in (4) with $\gamma = 1$.

2.2 Good-enough ancillary scaling

Since sub-Gaussian estimates of the variance $\mathbf{V}_\mu \mathbf{X}$ are not possible under our weak assumptions, it is natural to ask whether there exists a middle-ground, namely whether or not it is possible to construct a (data-driven) procedure for setting the scale $b > 0$ in (2) which is “good enough” in the sense that the resulting \mathbf{A}_n is sub-Gaussian, even though the scale itself cannot be. An initial (affirmative) answer to this question was given in recent work by Sun (2021), whose basic idea we briefly review here, with some slight re-formulation for readability and additional generality.

Essentially, the underlying idea in (Sun, 2021) is to utilize the convexity of ρ in (2), and to solve for both $a \in \mathbb{R}$ and $b > 0$ simultaneously, while penalizing b in such a way as to encourage scaling which is “good enough” as mentioned. More precisely, the empirical objective

$$\widehat{\mathcal{S}}_n(a, b) := \beta b + \frac{b}{n} \sum_{i=1}^n \rho\left(\frac{\mathbf{X}_i - a}{b}\right) \quad (5)$$

plays a central role, where $0 < \beta < 1$ is a parameter we can control, and ρ is fixed as

$$\rho(x) = \sqrt{x^2 + 1} - 1, \quad x \in \mathbb{R} \quad (6)$$

which is differentiable, and satisfies the Catoni condition (4) with $\gamma = 1$ (see Figure 1). If we fix $b > 0$, then the solution sets (in a) of both $\widehat{\mathcal{S}}_n(a, b)$ and $b \times \widehat{\mathcal{S}}_n(a, b)$ are identical, and it should be noted that the re-scaled map $x \mapsto b^2 \rho(x/b) = b\sqrt{x^2 + b^2} - b^2$ closely approximates $x \mapsto x^2/2$ as b grows large (Figure 2), and is well-known as the “pseudo Huber” or “smooth Huber” function, where b acts as a smoothing parameter.²

²Barron (2019, §1) gives a summary of this and related functions from the perspective of loss function design. This is not the only smoothed variant of the classic Huber function (Huber, 1964), see for example Rey (1983, §6.4.4).

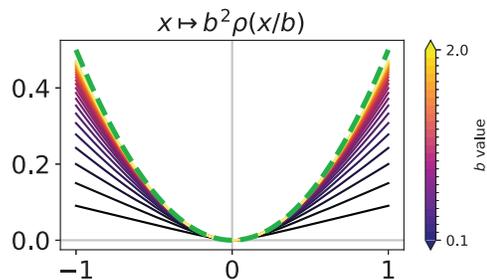


Figure 2: Graphs of the smooth Huber function, with ρ as in (6), over a range of smoothing parameters. For visual comparison, the graph of $x \mapsto x^2/2$ is plotted with a thick dashed green curve.

When considering the joint objective $\widehat{\mathcal{S}}_n(a, b)$, from the computational side, one important fact is that this function is convex on $\mathbb{R} \times (0, \infty)$ (see §A.2). From the statistical side of things, the solutions

$$(\mathbf{A}_n, \mathbf{B}_n) \in \arg \min_{a \in \mathbb{R}, b > 0} \widehat{\mathcal{S}}_n(a, b) \quad (7)$$

are such that under certain regularity conditions, the deviations $|\mathbf{A}_n - \mathbf{E}_\mu \mathbf{X}|$ are nearly optimal (sub-Gaussian, up to poly-logarithmic factors) (Sun, 2021, §3.3).³ The corresponding \mathbf{B}_n of course cannot give us sub-Gaussian estimates of the variance under such weak assumptions, but it does scale in a desirable way (Sun, 2021, §3.2), and when bias is mitigated by setting β sufficiently small given the sample size n , the resulting \mathbf{B}_n is good enough to provide such guarantees for \mathbf{A}_n , which is the ultimate goal anyways. By taking on a slightly more difficult optimization problem, it is possible to get away with not having prior knowledge or sub-Gaussian estimates of the variance. We use this basic insight as a stepping stone to our approach for learning algorithms charged with selecting a decision h such that the loss $\mathbf{L}(h)$ has a small mean-variance.

³Strictly speaking, the objective used in (Sun, 2021) is $\widehat{\mathcal{S}}_n(a, b)/\beta$, but all key results easily translate to our setup.

2.3 A bridge between two problems

To develop our proposal, we now return to the more general learning setup, where the test data is a random vector $Z \sim \mu$, test loss is $L(h) := \ell(h; Z)$, and we have n IID training points Z_1, \dots, Z_n yielding losses $L_i(\cdot) := \ell(\cdot; Z_i)$, $i \in [n]$. If our goal was to simply minimize the traditional risk $\mathbf{E}_\mu L(h)$ over $h \in \mathcal{H}$ under heavy-tailed losses, then in principle we could extend the approach of §2.2 to robustly estimate the test risk using

$$(\mathbf{A}_n(h), \mathbf{B}_n(h)) \in \arg \min_{a \in \mathbb{R}, b > 0} \left[\beta b + \frac{b}{n} \sum_{i=1}^n \rho \left(\frac{L_i(h) - a}{b} \right) \right] \quad (8)$$

and design a learning algorithm using (8) as follows:

$$\mathbf{H}_n \in \arg \min_{h \in \mathcal{H}} \mathbf{A}_n(h). \quad (9)$$

Under some regularity conditions, the machinery of Brownlees et al. (2015) could then be combined with pointwise concentration inequalities in (Sun, 2021) to control the tails of $\mathbf{E}_\mu L(\mathbf{H}_n)$ under just finite loss variance. Our goal however is not to minimize the expected loss, but rather the mean-SD sum (1). Furthermore, the bi-level program inherent in (9) is not computationally congenial from the perspective of large-scale machine learning tasks. To ease the computational burden while at the same time building a bridge between these two problems, we consider a new objective function taking the form

$$\widehat{\mathbf{C}}_n(h; a, b) := \alpha a + \beta b + \frac{\lambda b}{n} \sum_{i=1}^n \rho \left(\frac{L_i(h) - a}{b} \right) \quad (10)$$

with parameters $\alpha \geq 0$ and $\beta \geq 0$. We call (10) the *modified Sun-Huber objective*, since ρ is fixed as (6), and this form plays a special role in our analysis. Compared with that of (9), this objective is a simple function of h , and gradient-based minimizers can be easily applied assuming the underlying loss $\ell(\cdot)$ is sufficiently smooth. On the other hand, it is “biased” in the sense that it penalizes not just the loss location (whenever $\alpha > 0$), but the loss scale as well (whenever $\beta > 0$). Intuitively, some kind of deviation-driven “bias” is precisely what we need from the standpoint of minimizing the mean-SD objective $\text{MS}_\mu(h; \lambda)$, but it is not immediately clear how this objective relates to $\widehat{\mathbf{C}}_n(h; a, b)$, and it is equally unclear if we can just plug this new objective into standard machine learning workflows (e.g., using stochastic gradient-based optimizers) and achieve the desired effect without a prohibitive amount of tuning.

2.4 Overview of contributions and limitations

With our basic idea described and some key questions raised, we summarize the central points that charac-

terize the rest of this paper, and also highlight the limitations of this work. Broadly speaking, the new proposal here is a class of empirical “risk” minimizers, namely any learning algorithm which minimizes the new empirical objective (10). More explicitly, this refers to all procedures which returns a triplet satisfying

$$(\mathbf{H}_n, \mathbf{A}_n, \mathbf{B}_n) \in \arg \min_{h \in \mathcal{H}, a \in \mathbb{R}, b > 0} \widehat{\mathbf{C}}_n(h; a, b) \quad (11)$$

where \mathcal{H} denotes a set of feasible decisions, and we note that each element of this class is characterized by the settings of α , β , and λ used to define $\widehat{\mathbf{C}}_n$. In analogy with the strategy employed in §2.2, we do not expect \mathbf{A}_n and \mathbf{B}_n to provide sub-Gaussian estimates; we simply hope that these estimates are good enough to ensure the mean-SD is smaller and/or better-behaved when compared to standard benchmarks such as mean-based empirical risk minimization (ERM) and DRO-based algorithms. Theoretically, we are interested in identifying links between the proposed objective $\widehat{\mathbf{C}}_n$ and loss properties such as $\mathbf{E}_\mu L(h)$ and $\mathbf{V}_\mu L(h)$, with particular emphasis on how the settings of α , β , and λ influence such links.

Our main theory-driven contribution is the derivation of a principled approach to determine $\widehat{\mathbf{C}}_n$ (i.e., set α and β), before seeing any training data, in such a way that we can balance between “biased but robust” ρ -based deviations and “unbiased but outlier-sensitive” squared deviations that arise in the loss variance. Details are in §3.1–§3.3, and a concise procedure is summarized in Algorithm 1. We do not, however, consider the behavior of $\text{MS}_\mu(\mathbf{H}_n; \lambda)$ for a particular implementation of (11) (e.g., SGD) from a theoretical viewpoint; the implementation is left abstract. This is where the empirical analysis of §4 comes in. We provide evidence using simulated and real data that our procedure can be quite useful, even using a rudimentary implementation where we wrap base loss objects and naively pass them to standard stochastic gradient-based learning routines, with no manual tweaking of parameters.

3 THEORY

3.1 Links to the mean-SD objective

We would like to make the connection between the proposed objective (10) and the ultimate objective (1) a bit more transparent. To do this, we will make use of the population version of $\widehat{\mathbf{C}}_n$, denoted henceforth by C_μ and defined as

$$C_\mu(h; a, b) := \alpha a + \beta b + \lambda b \mathbf{E}_\mu \rho \left(\frac{L(h) - a}{b} \right). \quad (12)$$

Let us fix the decision h and threshold a , paying close attention to the optimal value of the scale b , denoted

here by $b_\mu(h, a)$. More explicitly, consider any positive real number satisfying

$$b_\mu(h, a) \in \arg \min_{b>0} C_\mu(h; a, b). \quad (13)$$

While it is not explicit in our notation, the optimal scale in (13) depends critically on the value of β . Intuitively, a smaller value of β leads to a weaker penalty for taking b large, thus encouraging a larger value of $b_\mu(h, a)$. In fact, one can show that viewing $b_\mu(h, a)$ as a function of the parameter β , in the limit we have

$$\lim_{\beta \rightarrow 0_+} b_\mu(h, a) = \infty. \quad (14)$$

See §C for a proof of (14). Combining this with the fact (also proved in §C) that

$$\lim_{b \rightarrow \infty} b \mathbf{E}_\mu \rho \left(\frac{\mathbf{L}(h) - a}{b} \right) = 0 \quad (15)$$

also holds, by re-scaling to avoid trivial limits we can obtain a result which sharply bounds the proposed learning criterion at the optimal scale using the square root of *quadratic* deviations, thereby establishing a clear link to the desired mean-SD objective (1).

Proposition 1. *Let \mathcal{H} be such that $\mathbf{E}_\mu |\mathbf{L}(h)|^2 < \infty$ for each $h \in \mathcal{H}$. If we set $\alpha = \alpha(\beta)$ such that $\alpha(\beta)/\sqrt{\beta} \rightarrow \tilde{\alpha} \in [0, \infty)$ as $\beta \rightarrow 0_+$, then in this limit, with appropriate re-scaling the scale-optimized learning criteria can be bounded above and below as*

$$\begin{aligned} & \tilde{\alpha} a + (1/2) \sqrt{\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} \\ & \leq \lim_{\beta \rightarrow 0_+} \min_{b>0} \frac{C_\mu(h; a, b)}{\sqrt{\beta}} \\ & \leq \tilde{\alpha} a + 4 \sqrt{\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} \end{aligned}$$

for any choice of threshold $a \in \mathbb{R}$ and weight $\alpha \geq 0$.

In the special case where $a = \mathbf{E}_\mu \mathbf{L}(h)$ and $\tilde{\alpha} > 0$, we naturally recover mean-SD sums akin to those studied in an ERM framework by Maurer and Pontil (2009) and those bounded from above using convex surrogates by Duchi and Namkoong (2019).

Of course in practice, we will only ever be working with fixed values of β , and the entire point of introducing new criteria (namely \hat{C}_n and C_μ) was to give us some control over how sensitive our objective is to loss tails. The following result makes the nature of this control (through β) more transparent.

Proposition 2. *Let \mathcal{H} and $\mathbf{L}(h)$ be as stated in Proposition 1. Letting $b_\mu(h, a)$ be as specified in (13), we define a Bernoulli random variable*

$$l(h; a) := \mathbf{I}\{|\mathbf{L}(h) - a| \leq b_\mu(h, a)\}$$

for any choice of $h \in \mathcal{H}$ and $a \in \mathbb{R}$. The optimal scale can then be bounded by

$$\frac{\lambda}{4\beta} \mathbf{E}_\mu l(h; a) (\mathbf{L}(h) - a)^2 \leq b_\mu^2(h, a) \leq \frac{\lambda}{2\beta} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2$$

for any choice of $0 < \beta < \lambda$ and $a \in \mathbb{R}$.

While it is difficult to pin down exactly how $b_\mu(h, a)$ changes as a function of β , Proposition 2 clearly shows us the appealing property that optimal scale induced by the proposed objective function essentially falls between the (tail-sensitive) quadratic deviations and a (tail-insensitive) truncated variant, with the truncation threshold loosening as β shrinks.

3.2 Guiding the optimal threshold

Since the preceding Propositions 1–2 both hold for any choice of threshold $a \in \mathbb{R}$, they clearly hold when both a and b are optimal, i.e., when a and b are set as

$$(a_\mu(h), b_\mu(h)) \in \arg \min_{a \in \mathbb{R}, b > 0} C_\mu(h; a, b). \quad (16)$$

In particular, using first-order conditions, the inclusion (16) is equivalent to the next two equalities holding:

$$\begin{aligned} \mathbf{E}_\mu \left(\frac{\mathbf{L}(h) - a_\mu(h)}{\sqrt{(\mathbf{L}(h) - a_\mu(h))^2 + b_\mu^2(h)}} \right) &= \alpha/\lambda, \\ \mathbf{E}_\mu \left(\frac{b_\mu(h)}{\sqrt{(\mathbf{L}(h) - a_\mu(h))^2 + b_\mu^2(h)}} \right) &= 1 - (\beta/\lambda). \end{aligned} \quad (17)$$

Given the context of our analysis in §3.1, let us consider the effect of taking β towards zero. For any non-trivial random loss, the second equality asks that $b_\mu(h)$ grow without bound as $\beta \rightarrow 0_+$, while $|a_\mu(h)|$ must be either bounded or grow slower than $b_\mu(h)$. On the other hand, if α is too large (i.e., $\alpha > \lambda$) then the first equality will be impossible to satisfy. In addition to taking $0 < \alpha < \lambda$, note that if we multiply both sides of the first equality in (17) by $b_\mu(h)$ and apply Proposition 2, then under this optimality condition we must have

$$\mathbf{E}_\mu \left(\frac{\mathbf{L}(h) - a_\mu(h)}{\sqrt{(\frac{\mathbf{L}(h) - a_\mu(h)}{b_\mu(h)})^2 + 1}} \right) \leq \alpha \sqrt{\frac{\lambda}{2\beta} \mathbf{E}_\mu (\mathbf{L}(h) - a_\mu(h))^2}. \quad (18)$$

With this inequality in place, we adopt the following strategy: encourage the optimal location to converge as $a_\mu(h) \rightarrow \mathbf{E}_\mu \mathbf{L}(h)$ when $\beta \rightarrow 0_+$. Since $\lambda > 0$ is assumed to be fixed in advance, the only way to ensure this using (18) is to set $\alpha = \alpha(\beta)$ such that

$$\lim_{\beta \rightarrow 0_+} \frac{\alpha(\beta)}{\sqrt{\beta}} = 0. \quad (19)$$

While (19) gives us a rather clear condition for determining α given β , we still do not have a principled setting for β . This point will be treated next.

3.3 Deriving an algorithm for finite samples

To complement the preceding analysis and discussion centered around the population objective (12), we now return to the empirical objective function $\widehat{\mathcal{C}}_n(h; a, b)$ introduced in (10). We maintain the running assumption that the training data Z_1, \dots, Z_n are an IID sample from μ , and thus the losses $L_i(h)$, $i = 1, \dots, n$ are independent given any fixed h . With h and $b > 0$ fixed for the moment, we will now take a closer look at the optimal (empirical) threshold that arises from this objective function, namely any random variable $A_n(h, b)$ satisfying

$$A_n(h, b) \in \arg \min_{a \in \mathbb{R}} \widehat{\mathcal{C}}_n(h; a, b). \quad (20)$$

Using the property (4) of the smooth Huber-like function ρ , we can demonstrate how data-driven thresholds satisfying (20) are concentrated at a point near the expected loss, where α and b play a key role in how close this point is to the mean.

Proposition 3 (Concentration at a shifted location). *Taking $0 \leq \alpha < 1$, $b > 0$, and $0 < \delta < 1$, with large enough n it is always possible to satisfy the condition*

$$\frac{4\alpha}{\lambda} \leq 4 \left(\frac{\mathbf{V}_\mu L(h)}{b^2} + \frac{\log(2/\delta)}{n} \right) \leq 1 - \frac{4\alpha}{\lambda},$$

and when this condition is satisfied, the data-driven threshold $A_n(h, b)$ in (20) satisfies

$$\begin{aligned} & \left| A_n(h, b) - \left[\mathbf{E}_\mu L(h) - \frac{2\alpha}{\lambda} b \right] \right| \\ & \leq 2 \left(\frac{\mathbf{V}_\mu L(h)}{b} + \frac{b \log(2/\delta)}{n} \right) \end{aligned}$$

with probability no less than $1 - \delta$.

This result can be seen as an extension of (Sun, 2021, Prop. 3.1) for the function (5) used in mean estimation to our generalized learning problem, although we use a different proof strategy which does not require strong convexity of $\widehat{\mathcal{C}}_n$ (with respect to a).

With Proposition 3 established, conventional wisdom might incline one to pursue a $\mathcal{O}(1/\sqrt{n})$ rate in the upper bound; in this case, setting $\beta \propto 1/n$ is a natural strategy since Proposition 2 tells us that for the population objective, the optimal setting of b scales with $\sqrt{\lambda/\beta}$. While this is natural from the perspective of tight concentration bounds for $A_n(h, b)$, we argue that a different strategy is more appropriate when we actually consider how (H_n, A_n, B_n) will behave in the

full joint optimization (11). The most obvious reason for this is that the joint objective lacks convexity and smoothness, as the following result summarizes.

Proposition 4 (Joint objective is non-convex and non-smooth). *Even when \mathcal{H} is a compact convex set and the base loss function $\ell(\cdot; Z)$ is convex, the mapping $(h, a, b) \mapsto \widehat{\mathcal{C}}_n$ is not convex in general, and is non-smooth in the sense that its gradient is not Lipschitz continuous on $\mathcal{H} \times \mathbb{R} \times (0, \infty)$.*

In consideration of Proposition 4, standard complexity results for typical optimizers such as stochastic gradient descent to achieve a ε -stationary point are on the order of $\mathcal{O}(\varepsilon^{-4})$; see Davis and Drusvyatskiy (2019) for example.⁴ With this in mind, setting $\beta \propto 1/n$ to achieve $\mathcal{O}(\varepsilon^{-2})$ sample complexity for error bounds of $A_n(h, b)$ seems superfluous if in the end the dominant complexity for solving the ultimate problem (11) will be of the order $\mathcal{O}(\varepsilon^{-4})$. As such, in order to match this rate, the more natural strategy is to set $\beta \propto 1/\sqrt{n}$, or more precisely to set

$$\beta = \frac{\beta_0}{\sqrt{n}} \quad (21)$$

where $\beta_0 > 0$ is a constant used to ensure $0 < \beta < \lambda$. This, coupled with $\alpha(\beta) = \beta$ to satisfy (19) from the previous sub-section, is our proposed setting to determine (α, β) (and thus $\widehat{\mathcal{C}}_n$) using just knowledge of n , and without having observed any data points. This procedure is summarized in Algorithm 1, and will be studied empirically in §4.

Algorithm 1 Modified Sun-Huber

Inputs: data Z_1, \dots, Z_n and parameter $\lambda > 0$.

Set: $\beta = \beta_0/\sqrt{n}$, with β_0 such that $0 < \beta < \lambda$.

Set: $\alpha = \beta$.

Minimize: $\widehat{\mathcal{C}}_n(h; a, b)$, with α and β as above.

4 EMPIRICAL ANALYSIS

Our investigation in the previous section led us to Algorithm 1, giving us a principled and precise strategy to construct the objective function $\widehat{\mathcal{C}}_n$, but leaving the actual minimization procedure abstract. Here we make this concrete by implementing a simple gradient-based minimizer of this objective, and comparing this procedure with natural benchmarks from the literature.

4.1 Methods to be compared

In the experiments to follow in §4.2–§4.3, we compare our proposed procedure (denoted in figures as “Modified

⁴Even if the objective were smooth, the same rates are typical; see for example Ghadimi and Lan (2013).

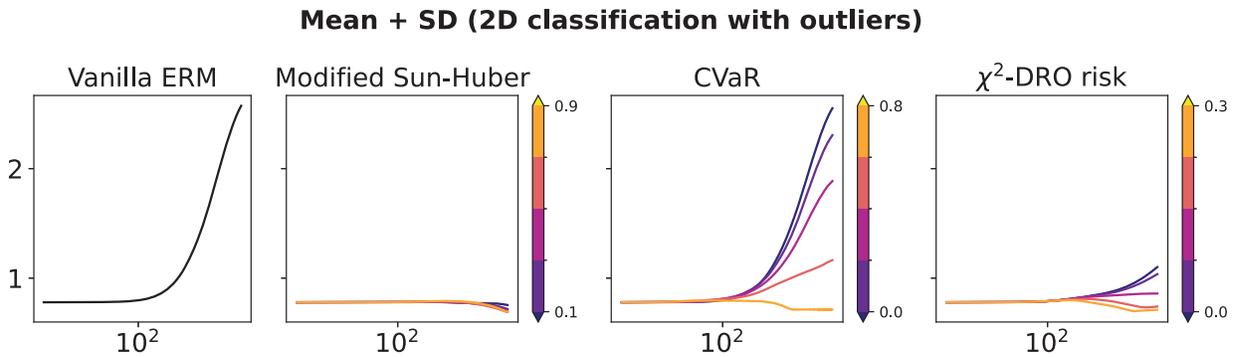


Figure 3: Trajectory of the (empirical) mean-SD objective (1) over iterations. Colors correspond to different choices from each class: β_0 for Modified Sun-Huber, quantile level for CVaR, and robustness level for DRO.

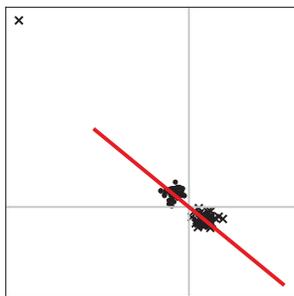


Figure 4: 2D classification example from §4.2. The red line represents the initial value used by each method.

Sun-Huber”) with three alternatives: traditional mean-based empirical risk minimization (denoted “Vanilla ERM”), conditional value-at-risk (CVaR) (Curi et al., 2020), and the well-studied χ^2 -DRO risk (Duchi and Namkoong, 2019; Hashimoto et al., 2018). For reference, here we provide the population versions of the CVaR and χ^2 -DRO criteria used in the empirical tests to follow. First, it is well known (see Rockafellar and Uryasev (2000)) that CVaR at quantile level ξ can be represented as

$$\text{CVaR}_\mu(h; \xi) = \inf_{a \in \mathbb{R}} \left[a + \frac{1}{1 - \xi} \mathbf{E}_\mu (L(h) - a)_+ \right] \quad (22)$$

where $(x)_+ := \max\{0, x\}$. Similarly, DRO risk based on the Cressie-Read family of divergence functions, here denoted by $\text{DRO}_\mu(h; \eta)$, is formulated for any $c > 1$ and $\eta > 0$ using

$$\inf_{a \in \mathbb{R}} \left[a + (1 + c(c - 1)\eta)^{1/c} (\mathbf{E}_\mu (L(h) - a)_+)^{1/c_*} \right] \quad (23)$$

where $c_* := c/(c - 1)$, and χ^2 -DRO is the special case where $c = 2$ (Duchi and Namkoong, 2019; Hashimoto et al., 2018; Zhai et al., 2021). The different “robustness

Trajectory with best final error rate

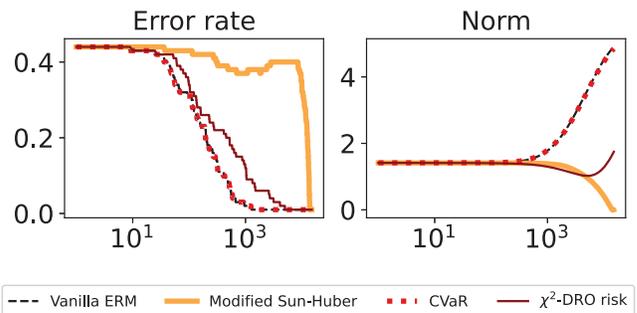


Figure 5: From each method class, we show the classification error rate and Euclidean norm trajectories corresponding to the setting that achieved the best error rate after the final iteration.

levels” to be mentioned in §4.2 correspond to different values of the re-parameterized quantity $\tilde{\eta} \in (0, 1)$, related to η by the equality $\eta = (1/(1 - \tilde{\eta}) - 1)/2$. Just as our $\hat{C}_n(h; a, b)$ is solved jointly in (h, a, b) , our empirical tests minimize the empirical versions of (22) and (23) jointly in (h, a) .

4.2 Simulated noisy classification on the plane

As a simplified and controlled setting to start with, we generate random data points on the plane which are mostly linearly separable, save for a single distant outlier (Figure 4). Before we consider off-sample generalization, here we focus simply on the training loss distribution properties as a function of algorithm iterations.

Experiment setup We generate $n = 100$ training data points using two Gaussian distributions on the plane to represent two classes, with each class having the same number of points. We choose a single point

uniformly at random, and perturb it by multiplying the scalar -10. As mentioned in §4.1, we compare our proposed procedure (Modified Sun-Huber) with Vanilla ERM, CVaR, and χ^2 -DRO. In light of Algorithm 1, we set $\lambda = \log(n)/\sqrt{n} > \beta = \beta_0/\sqrt{n}$, and try a variety of β_0 values just for reference. For all the aforementioned methods, we set the base loss $\ell(\cdot)$ to the usual binary logistic loss (linear model), and run (batch) gradient descent on the empirical risk objectives implied by each of these methods, with a fixed step size of 0.01 over 15,000 iterations. Alternative settings of step size and iteration number were not tested. All methods are initialized at the same point, shown in Figure 4.

Results and discussion In Figure 3, we show the empirical mean-SD trajectories for the base loss, over algorithm iterations (\log_{10} scale), for each method of interest. Using our notation, this is the sample version of $MS_\mu(h; 1)$ in (1); note that this differs from the n -dependent λ setting that is explicit in our proposed method, and implicit in CVaR and χ^2 -DRO. All methods besides vanilla ERM have multiple settings that were tested, and the results for each are distinguished using curves of different color. Our method tests different values of β_0 , CVaR tests different quantile levels, and DRO tests different robustness levels. Since Vanilla ERM is designed to optimize the average loss, it is perhaps not surprising that it fails in terms of the mean-SD objective. On the other hand, the proposed method (for any choice of β_0) is as good or better than all the competing methods. As a basic sanity check, in Figure 5 we also consider the error rate (average zero-one loss) and model norm trajectories over iterations for each method. For each method, we plot just one trajectory, namely the one achieving the best final error rate. While our method is not designed to minimize the average loss and typical surrogate theory does not apply, the error rate is surprisingly good, albeit with slower convergence than the other methods. The error rate for CVaR matches that of Vanilla ERM; this is in fact the CVaR setting with the worst final mean-SD value. On the other hand, the proposed method performs well from both perspectives at once.

4.3 Classification on real datasets

We proceed to experiments using real-world datasets, some of which are orders of magnitude larger than the simple setup given in §4.2, and which include multi-class classification tasks.

Experiment setup We make use of four well-known datasets, all available from online repositories: `adult`, `australian`, `cifar10`, and `fashion_mnist`. For multi-class datasets, we extend the binary logistic loss to the usual multi-class logistic regression loss under a

linear model, with one linear model for each class. Features for all datasets are normalized to $[0, 1]$, with one-hot representations of categorical features. The learning algorithms being compared here are the same as described in §4.2, except that now we implement each method using mini-batch stochastic gradient descent (batch size 32), and do 30 epochs (i.e., 30 passes over the training data). In addition, our proposed “Modified Sun-Huber” method performs almost identically for the range of β_0 values tested in §4.2, and thus we have simply fixed $\beta_0 = 0.9$, so there is only one trajectory curve this time. On the other hand, we now try a range of step sizes for each method, choosing the best step size in terms of average (base) loss value on validation data for each method. We run five independent trials, and for each trial we randomly re-shuffle the dataset, taking 80% for training, 10% for validation (used to select step sizes), and 10% for final testing.

Results and discussion Our main results are shown in Figure 6 (next page), where once again we plot the trajectory of the mean-SD objective, but this time computed on test data, and given as a function of *epoch number*, rather than individual iterations. Curves drawn represent averages taken over trials, and the lightly shaded region above/below each curve shows standard deviation over trials. Perhaps surprisingly, the very simple implementation of our proposed Algorithm 1 (fixed step size, no regularization) works remarkably well on a number of datasets. From the perspective of mean-SD minimization, for three out of four datasets, the proposed method is far better than Vanilla ERM, and as good or better than even the best settings of CVaR and DRO viewed after the fact. Regarding the sub-standard performance observed on `fashion_mnist`, detailed analysis shows that more fine-tuned settings of α and β can readily bring the method up to par; the non-convex and non-smooth nature of \hat{C}_n naturally means that some tasks will require more careful settings than are captured by our Algorithm 1, and indeed will take explicit account of the optimizer to be used. We leave both the theoretical grounding and empirical testing of such optimizer-aligned mean-SD minimizers for future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H03646. Thanks also go to anonymous reviewers whose careful reading and insightful comments contributed to the final version of this paper.

References

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

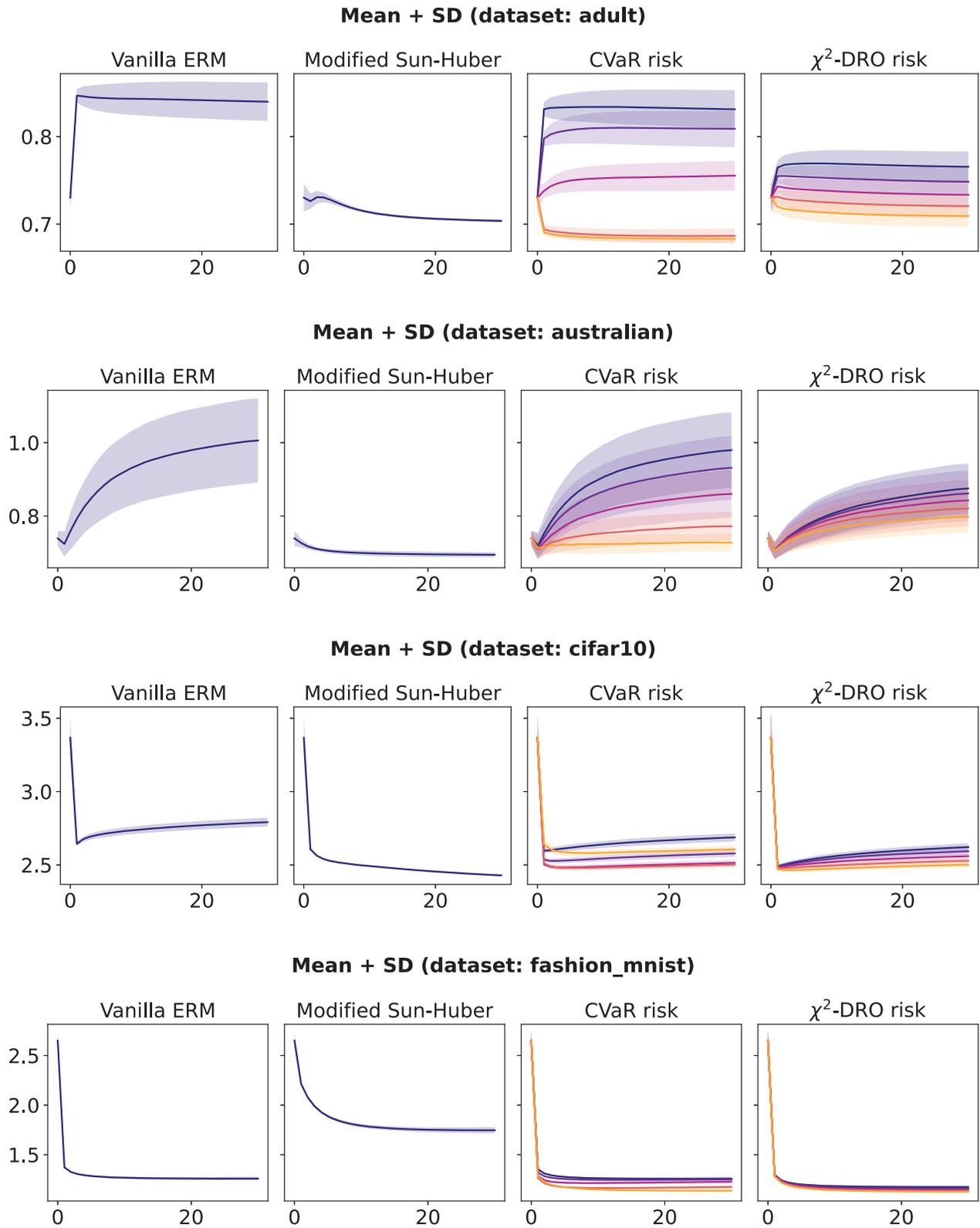


Figure 6: Mean-SD trajectories on real-world datasets as described in §4.3, given as a function of epochs and averaged over multiple independent trials. Coloring for CVaR and DRO is analogous to that of Figure 3.

- Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339.
- Ben-Tal, A., Den Hertog, D., De Waegenare, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
- Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536.
- 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)*, pages 1036–1047.
- Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239.
- Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
- Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.
- Duchi, J. and Namkoong, H. (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55.
- Ghadimi, S. and Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.
- Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938.
- Holland, M. J. and Tanabe, K. (2023). A survey of learning criteria going beyond the usual risk. *Journal of Artificial Intelligence Research*, 73:781–821.
- Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40.
- Hu, S., Wang, X., and Lyu, S. (2022). Rank-based decomposable losses in machine learning: A survey. *arXiv preprint arXiv:2207.08768v1*.
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, 2nd edition.
- 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.
- Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2021). Tilted empirical risk minimization. In *The 9th International Conference on Learning Representations (ICLR)*.
- Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.
- Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.
- Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*.
- Medina, A. M. and Yang, S. (2021). Robust unsupervised learning via L-statistic minimization. In *38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 7524–7533.
- Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2021). Long-tail learning via logit adjustment. In *The 9th International Conference on Learning Representations (ICLR)*.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.
- Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.
- Rey, W. J. J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer.
- Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42.
- Rockafellar, R. T. and Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53.
- Royset, J. O. (2022). Risk-adaptive approaches to learning and decision making: A survey. *arXiv preprint arXiv:2212.00856*.

- Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275.
- Sun, Q. (2021). Do we need to estimate the variance in robust mean estimation? *arXiv preprint arXiv:2107.00118v1*.
- Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2nd edition.
- Zhai, R., Dan, C., Kolter, J. Z., and Ravikumar, P. (2021). DORO: Distributional and outlier robust optimization. In *38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12345–12355.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Source code is public on GitHub.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes]
 - (b) The license information of the assets, if applicable. [Not Applicable]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Source code is made available at a public GitHub repository.
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A TECHNICAL APPENDIX

A.1 Basic facts

Assuming ρ is defined as in (6), let us consider the function

$$f(x, a, b) := \alpha a + \beta b + b\rho\left(\frac{x-a}{b}\right) \quad (24)$$

$$= \alpha a + \beta b + \sqrt{(x-a)^2 + b^2} - b \quad (25)$$

$$= \alpha a + \sqrt{(x-a)^2 + b^2} - (1-\beta)b. \quad (26)$$

The partial derivatives are as follows.

$$\partial_x f(x, a, b) = \frac{x-a}{\sqrt{(x-a)^2 + b^2}} \quad (27)$$

$$\partial_a f(x, a, b) = \alpha - \frac{x-a}{\sqrt{(x-a)^2 + b^2}} \quad (28)$$

$$\partial_b f(x, a, b) = \frac{b}{\sqrt{(x-a)^2 + b^2}} - (1-\beta) \quad (29)$$

The corresponding second derivatives are as follows.

$$\partial_x^2 f(x, a, b) = \frac{1}{\sqrt{(x-a)^2 + b^2}} - \frac{(x-a)^2}{((x-a)^2 + b^2)^{3/2}} = \frac{b^2}{((x-a)^2 + b^2)^{3/2}} \quad (30)$$

$$\partial_a^2 f(x, a, b) = \frac{1}{\sqrt{(x-a)^2 + b^2}} - \frac{(x-a)^2}{((x-a)^2 + b^2)^{3/2}} = \frac{b^2}{((x-a)^2 + b^2)^{3/2}} \quad (31)$$

$$\partial_b^2 f(x, a, b) = \frac{1}{\sqrt{(x-a)^2 + b^2}} - \frac{b^2}{((x-a)^2 + b^2)^{3/2}} = \frac{(x-a)^2}{((x-a)^2 + b^2)^{3/2}} \quad (32)$$

The remaining elements of the Hessian of $f(x, a, b)$ follow easily, given as follows.

$$\partial_a \partial_x f(x, a, b) = \frac{-1}{\sqrt{(x-a)^2 + b^2}} + \frac{(x-a)^2}{((x-a)^2 + b^2)^{3/2}} = \frac{-b^2}{((x-a)^2 + b^2)^{3/2}} \quad (33)$$

$$\partial_b \partial_x f(x, a, b) = \frac{-b(x-a)}{((x-a)^2 + b^2)^{3/2}} \quad (34)$$

$$\partial_b \partial_a f(x, a, b) = \frac{b(x-a)}{((x-a)^2 + b^2)^{3/2}} \quad (35)$$

Lemma 5 (Useful inequalities).

$$\frac{1}{1+x} \leq 1 - \frac{x}{2}, \quad 0 \leq x \leq 1. \quad (36)$$

$$(1+x)^c \geq 1+cx, \quad x \geq -1, c \in \mathbb{R} \setminus (0, 1). \quad (37)$$

A.2 Convexity and smoothness

Lemma 6. *The map $x \mapsto 1/\sqrt{1+x}$ is convex on $[0, \infty)$.*

Lemma 7 (Properties of partial objective). *With ρ as in (6) and $\beta \geq 0$, the function*

$$(x, b) \mapsto \beta b + b\rho\left(\frac{x}{b}\right)$$

is convex and $(1 + \max\{1 - \beta, \beta\})$ -Lipschitz (in $\|\cdot\|_1$) on $\mathbb{R} \times (0, \infty)$, but its gradient is not (globally) Lipschitz, and thus the function is not smooth.⁵

Proof of Lemma 7. For notational convenience, setting $0 < \beta < 1$, let us denote

$$g(x, b) := \beta b + b\rho(x/b), \quad x \in \mathbb{R}, b > 0$$

with ρ as in (6). From the partial derivatives (27) and (29), it is clear that we have

$$-1 \leq \partial_x g(x, b) \leq 1, \quad -(1 - \beta) \leq \partial_b g(x, b) \leq \beta$$

when evaluated at any choice of $x \in \mathbb{R}$ and $b > 0$. It follows that the gradient norm can be bounded as

$$\|\nabla g(x, b)\|_1 \leq 1 + \max\{(1 - \beta), \beta\}$$

and thus $g(\cdot)$ is Lipschitz continuous in $\|\cdot\|_1$ (and also $\|\cdot\|_2$).⁶

Next, let us denote the Hessian of $g(\cdot)$ evaluated at (x, b) by H . Basic calculus gives us the simple form

$$H := \frac{1}{(x^2 + b^2)^{3/2}} \begin{bmatrix} b^2 & -xb \\ -xb & x^2 \end{bmatrix}$$

and for any pair of real values $\mathbf{u} = (u_1, u_2)$, we have

$$\langle H\mathbf{u}, \mathbf{u} \rangle = \frac{1}{(x^2 + b^2)^{3/2}} (u_1 b - u_2 x)^2 \geq 0. \tag{38}$$

Since this holds for any choice of $x \in \mathbb{R}$ and $b > 0$, the Hessian is thus positive semi-definite, implying that $g(\cdot)$ is (jointly) convex (Nesterov, 2004, Thm. 2.1.4).

On the other hand, the function $g(\cdot)$ is not smooth. To see this, first note that having chosen any \mathbf{u} such that $\|\mathbf{u}\| \leq 1$, we have that the (operator) norm is bounded below as

$$\|H\| = \sup_{\|\mathbf{u}'\| \leq 1} \left[\sup_{\|\mathbf{u}''\| \leq 1} \langle H\mathbf{u}', \mathbf{u}'' \rangle \right] \geq \langle H\mathbf{u}, \mathbf{u} \rangle.$$

Then, as a concrete example, consider setting $x = b$, with $\mathbf{u} = (u_1, u_2)$ such that $u_1 \neq u_2$. Recalling the lower bound (38), we have

$$\|H\| \geq \frac{b^2}{(2b^2)^{3/2}} (u_1 - u_2)^2 = \frac{(u_1 - u_2)^2}{(\sqrt{2})^3 b} \rightarrow \infty$$

in the limit as $b \rightarrow 0_+$. As such, the gradient of $g(\cdot)$ cannot be Lipschitz continuous on $\mathbb{R} \times (0, \infty)$, and thus $g(\cdot)$ is not smooth (Nesterov, 2004, Thm. 2.1.6). \square

⁵We prove that the Hessian's norm is unbounded, which implies (via Nesterov (2004, Thm. 2.1.6)) that the convex function of interest cannot be smooth.

⁶That bounded gradients imply Lipschitz continuity is a general fact on linear spaces (Luenberger, 1969, §7.3, Prop. 2).

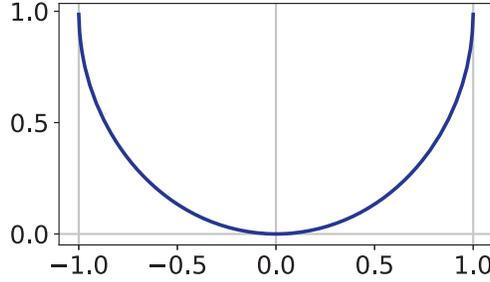


Figure 7: Graph of the Legendre transform ρ^* as given in (40) over $(-1, 1)$.

B APPENDIX TO THE MAIN BODY

B.1 Comparison with dual form of DRO risk

Some readers may notice that the proposed (population) objective (12) looks quite similar to the dual form of DRO risks:

$$\text{DRO}_\mu(h; \beta) := \inf_{a \in \mathbb{R}, b > 0} \left[a + \beta b + b \mathbf{E}_\mu \phi^* \left(\frac{\mathbf{L}(h) - a}{b} \right) \right] \quad (39)$$

where ϕ^* is the Legendre-Fenchel convex conjugate $\phi^*(x) := \sup_{u \in \mathbb{R}} [xu - \phi(u)]$ induced by a function $\phi : \mathbb{R} \rightarrow \bar{\mathbb{R}}$, assumed to be convex and lower semi-continuous, with $\phi(1) = 0$ and $\phi(x) = \infty$ whenever $x < 0$; see Shapiro (2017, §3.2) for reference. Given this similarity, one might ask whether or not some form of DRO risk can be reverse engineered from our proposed objective. Taking up this point briefly, we first note that the conjugate of ρ given by (6) is

$$\rho^*(x) := \sup_{u \in \mathbb{R}} [xu - \rho(u)] = \sup_{u \in \mathbb{R}} [xu - \sqrt{u^2 + 1} + 1].$$

From the non-negative nature of ρ , clearly $\rho^*(0) = -\rho(0) = 0$. For $x \neq 0$, note that taking the derivative of concave function $u \mapsto xu - \rho(u)$ and setting it to zero, we obtain the first-order optimality conditions

$$\frac{u}{\sqrt{u^2 + 1}} = x \iff \frac{\text{sign}(x)}{\sqrt{1 + 1/u^2}} = x \iff \frac{1}{x^2} = 1 + 1/u^2 \iff u = \frac{\text{sign}(x)}{\sqrt{1/x^2 - 1}}.$$

Plugging this solution in whenever $|x| < 1$ and doing a bit of algebra readily yields the simple closed-form expression

$$\rho^*(x) = \begin{cases} \frac{x^2}{\sqrt{1-x^2}} + 1 - \frac{1}{\sqrt{1-x^2}}, & \text{if } 0 \leq |x| < 1 \\ \infty, & \text{else.} \end{cases} \quad (40)$$

As can be readily observed from both (40) and Figure 7, this function does not satisfy any of the requirements placed on ϕ except convexity, and thus despite the similar form, the non-monotonic nature of ρ is in sharp contrast with monotonicity of typical cases of ϕ^* that arise in the DRO literature, e.g. Ben-Tal et al. (2013, §3), and does not readily imply a “primal” DRO objective that can be recovered using ρ^* .

B.2 Stationary points of mean-variance

Having established links between the proposed objective and the mean-SD objective, we next consider the mean-variance objective

$$\text{MV}_\mu(h) := \mathbf{E}_\mu \mathbf{L}(h) + \mathbf{V}_\mu \mathbf{L}(h). \quad (41)$$

This quantity can be expressed as the minimum value of a convex function, namely we have

$$\text{MV}_\mu(h) = \min_{a \in \mathbb{R}} \left[a + \frac{\mathbf{E}_\mu(\mathbf{L}(h) - a)^2 + 1}{2} \right] = a_{\text{MV}}(h) + \frac{\mathbf{E}_\mu(\mathbf{L}(h) - a_{\text{MV}}(h))^2 + 1}{2} \quad (42)$$

where on the right-most side we have set $a_{\text{MV}}(h) := \mathbf{E}_\mu \mathbf{L}(h) - 1$. Assuming the underlying loss is differentiable, the gradient with respect to h can be written as

$$\begin{aligned} \text{MV}'_\mu(h) &= \mathbf{E}_\mu \mathbf{L}'(h) + \mathbf{E}_\mu \mathbf{L}(h) \mathbf{L}'(h) - \mathbf{E}_\mu \mathbf{L}(h) \mathbf{E}_\mu \mathbf{L}'(h) \\ &= \mathbf{E}_\mu \mathbf{L}'(h) + \mathbf{E}_\mu (\mathbf{L}(h) - \mathbf{E}_\mu \mathbf{L}(h)) \mathbf{L}'(h) \\ &= \mathbf{E}_\mu (\mathbf{L}(h) - (\mathbf{E}_\mu \mathbf{L}(h) - 1)) \mathbf{L}'(h) \end{aligned}$$

which implies a stationarity condition of

$$\text{MV}'_\mu(h) = 0 \iff \mathbf{E}_\mu (\mathbf{L}(h) - (\mathbf{E}_\mu \mathbf{L}(h) - 1)) \mathbf{L}'(h) = 0. \quad (43)$$

Similarly, the partial derivative of the learning criterion (12) taken with respect to h is

$$\frac{\partial}{\partial h} C_\mu(h; a, b) = \mathbf{E}_\mu \left(\frac{\mathbf{L}(h) - a}{\sqrt{(\mathbf{L}(h) - a)^2 + b^2}} \right) \mathbf{L}'(h)$$

and thus multiplying both sides by $b > 0$, we obtain a simple stationarity condition of

$$\frac{\partial}{\partial h} C_\mu(h; a, b) = 0 \iff \mathbf{E}_\mu \left(\frac{\mathbf{L}(h) - a}{\sqrt{(\frac{\mathbf{L}(h) - a}{b})^2 + 1}} \right) \mathbf{L}'(h) = 0. \quad (44)$$

With the right threshold setting, obviously the two conditions become very similar as b grows large. The following result makes this precise.

Proposition 8. *Let loss function ℓ and data distribution μ be such that the random vector $\mathbf{L}(h) \mathbf{L}'(h)$ is integrable and has a norm with finite mean, i.e., $\mathbf{E}_\mu \|\mathbf{L}(h) \mathbf{L}'(h)\| < \infty$ for some choice of $h \in \mathcal{H}$. Then, for any $a \in \mathbb{R}$, defining*

$$f(h; a) := \lim_{b \rightarrow \infty} b \frac{\partial}{\partial h} C_\mu(h; a, b) \quad (45)$$

the stationary points of the mean-variance objective are related to those of the proposed objective (12) through the following equivalence:

$$f(h; a_{\text{MV}}(h)) = 0 \iff \frac{\partial}{\partial h} \text{MV}_\mu(h) = 0$$

where $\text{MV}_\mu(h)$ is as defined in (41).

C ADDITIONAL PROOFS

Proof of Proposition 1. To begin, note that the function

$$b \mapsto b \mathbf{E}_\mu \rho \left(\frac{\mathbf{L}(h) - a}{b} \right) = \mathbf{E}_\mu \left[\sqrt{(\mathbf{L}(h) - a)^2 + b^2} - b \right] \quad (46)$$

is monotonic (non-increasing) on $(0, \infty)$ (follows clearly from (29)). We will use this property moving forward. Recalling the upper and lower bounds of Proposition 2, we re-write them as

$$\frac{c_{\text{lo}}(\beta)}{\beta} \leq b_\mu^2(h, a) \leq \frac{c_{\text{hi}}}{\beta} \quad (47)$$

using the shorthand notation

$$\begin{aligned} c_{\text{lo}}(\beta) &:= \frac{\lambda}{4} \mathbf{E}_\mu \mathbf{l}(h; a) (\mathbf{L}(h) - a)^2 \\ c_{\text{hi}} &:= \frac{\lambda}{2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2 \end{aligned}$$

and noting that while c_{hi} is free of β , $c_{\text{lo}}(\beta)$ depends on β through the definition of $\mathbf{l}(h; a)$. Fixing $0 < \beta < \lambda$ for now and recalling the form of C_μ in (12), the preceding bounds (47) and monotonicity of (46) can be used to obtain a lower bound of the form

$$\min_{b>0} C_\mu(h; a, b) \geq \alpha a + \sqrt{\beta c_{\text{lo}}(\beta)} + \lambda \sqrt{c_{\text{hi}}} \mathbf{E}_\mu \left[\sqrt{\frac{(\mathbf{L}(h) - a)^2}{c_{\text{hi}}} + \frac{1}{\beta}} - \sqrt{\frac{1}{\beta}} \right]. \quad (48)$$

Using the fact (14) and applying dominated convergence (Ash and Doléans-Dade, 2000, Thm. 1.6.9), in the limit we have

$$\lim_{\beta \rightarrow 0_+} c_{\text{lo}}(\beta) = \frac{\lambda}{4} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2.$$

Dividing both sides of (48) by $\sqrt{\beta}$, setting $\alpha = \alpha(\beta)$ as in the proposition statement, and taking the limit as $\beta \rightarrow 0_+$, we obtain

$$\begin{aligned} \lim_{\beta \rightarrow 0_+} \min_{b>0} \frac{C_\mu(h; a, b)}{\sqrt{\beta}} &\geq \tilde{\alpha} a + \sqrt{\frac{\lambda}{4} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} + \frac{\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2}{2\sqrt{c_{\text{hi}}}} \\ &= \tilde{\alpha} a + \sqrt{\frac{\lambda}{4} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} + \sqrt{\frac{\lambda}{2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} \\ &= \tilde{\alpha} a + \left(\frac{1}{2} + \frac{1}{\sqrt{2}} \right) \sqrt{\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2}. \end{aligned}$$

The first inequality uses the fact that for any $c > 0$, we have $\sqrt{cx + x^2} - x \rightarrow c/2$ as $x \rightarrow \infty$, and also uses dominated convergence. The remaining equalities just follow from plugging in the definition of c_{hi} and cleaning up terms. This proves the desired lower bound.

As for the upper bound of interest, a perfectly analogous argument can be applied. Using Proposition 2 again and taking β small enough that

$$c_{\text{lo}}(\beta) \geq c_{\text{hi}}/4 \quad (49)$$

holds (always possible), we can obtain upper bounds of the form

$$\begin{aligned} \min_{b>0} C_\mu(h; a, b) &\leq \alpha a + \sqrt{\beta c_{\text{hi}}} + \lambda \sqrt{c_{\text{lo}}(\beta)} \mathbf{E}_\mu \left[\sqrt{\frac{(\mathbf{L}(h) - a)^2}{c_{\text{lo}}(\beta)} + \frac{1}{\beta}} - \sqrt{\frac{1}{\beta}} \right] \\ &\leq \alpha a + \sqrt{\beta c_{\text{hi}}} + \lambda \sqrt{c_{\text{hi}}} \mathbf{E}_\mu \left[\sqrt{\frac{4(\mathbf{L}(h) - a)^2}{c_{\text{hi}}} + \frac{1}{\beta}} - \sqrt{\frac{1}{\beta}} \right] \end{aligned} \quad (50)$$

noting that the latter inequality (50) follows from using (49) as well as $c_{\text{lo}}(\beta) \leq c_{\text{hi}}$. As with the lower bound argument in the preceding paragraph, we set $\alpha = \alpha(\beta)$, divide both sides by $\sqrt{\beta}$, and take the limit as $\beta \rightarrow 0_+$. This results in

$$\begin{aligned} \lim_{\beta \rightarrow 0_+} \min_{b>0} \frac{C_\mu(h; a, b)}{\sqrt{\beta}} &\leq \tilde{\alpha} a + \sqrt{\frac{\lambda}{2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} + \frac{2\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2}{\sqrt{c_{\text{hi}}}} \\ &= \tilde{\alpha} a + \sqrt{\frac{\lambda}{2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} + 2\sqrt{2\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} \\ &= \tilde{\alpha} a + \left(2\sqrt{2} + \frac{1}{\sqrt{2}} \right) \sqrt{\lambda \mathbf{E}_\mu (\mathbf{L}(h) - a)^2} \end{aligned}$$

which gives us the desired upper bound. The bounds given in the proposition statement are slightly looser, but more readable. \square

Proof of Proposition 2. We adapt key elements of the scale control used by Sun (2021, §2) to our setting. We start by looking at first-order conditions for optimality of $b > 0$. First, note that

$$\begin{aligned} \frac{\partial}{\partial b} C_\mu(h; a, b) &= \beta + \lambda \frac{\partial}{\partial b} \left(\mathbf{E}_\mu \sqrt{(\mathbf{L}(h) - a)^2 + b^2} - b \right) \\ &= \beta + \lambda \mathbf{E}_\mu \left(\frac{b}{\sqrt{(\mathbf{L}(h) - a)^2 + b^2}} \right) - \lambda. \end{aligned}$$

As such, it follows that

$$\mathbf{E}_\mu \left(\frac{b}{\sqrt{(\mathbf{L}(h) - a)^2 + b^2}} \right) = 1 - \beta/\lambda \tag{51}$$

is equivalent to $\partial_b C_\mu(h; a, b) = 0$. Obviously, the left-hand side of (51) is non-negative for all $b \geq 0$ and bounded above by 1 for all $b \geq 0$, $a \in \mathbb{R}$, and $h \in \mathcal{H}$. Thus (51) can only hold for $0 \leq \beta \leq \lambda$. Using convexity (Lemma 6) and Jensen's inequality (Ash and Doléans-Dade, 2000, Thm. 6.3.5), we have

$$\mathbf{E}_\mu \left(\frac{b}{\sqrt{(\mathbf{L}(h) - a)^2 + b^2}} \right) = \mathbf{E}_\mu \left(\frac{1}{\sqrt{(\frac{\mathbf{L}(h) - a}{b})^2 + 1}} \right) \geq \left(\frac{1}{\sqrt{\mathbf{E}_\mu \left(\frac{(\mathbf{L}(h) - a)^2}{b^2} + 1 \right)}} \right)$$

and thus whenever (51) holds, we know that

$$(1 - \beta/\lambda)^2 \geq \frac{1}{\mathbf{E}_\mu \left(\frac{(\mathbf{L}(h) - a)^2}{b^2} + 1 \right)}$$

must also hold. Re-arranging terms, we see that this implies

$$b^2 \leq \frac{(1 - \beta/\lambda)^2 \mathbf{E}_\mu (\mathbf{L}(h) - a)^2}{1 - (1 - \beta/\lambda)^2}.$$

For readability, set $\eta := \beta/\lambda$, and note that since

$$\frac{(1 - \eta)^2}{1 - (1 - \eta)^2} = \frac{(1 - \eta)^2}{2\eta - \eta^2} = \frac{(1 - \eta)^2}{2\eta(1 - \eta/2)} \leq \frac{(1 - \eta)^2}{2\eta(1 - \eta)} \leq \frac{1}{2\eta}$$

we can obtain the cleaner (but looser) upper bound

$$b^2 \leq \frac{\mathbf{E}_\mu (\mathbf{L}(h) - a)^2}{2\eta} = \frac{\lambda}{2\beta} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2$$

for any choice of $0 < \beta \leq \lambda$ and $a \in \mathbb{R}$. Since the first-order condition (51) is necessary for optimality (Nesterov, 2004, Thm. 1.2.1), it follows that

$$b_\mu^2(h, a) \leq \frac{\lambda}{2\beta} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2 \tag{52}$$

which is the desired upper bound.

Considering a lower bound next, note first that using the concavity of $x \mapsto \sqrt{x}$ on \mathbb{R}_+ , another application of Jensen's inequality gives us

$$\mathbf{E}_\mu \left(\frac{b}{\sqrt{(\mathbf{L}(h) - a)^2 + b^2}} \right) = \mathbf{E}_\mu \sqrt{\frac{b^2}{(\mathbf{L}(h) - a)^2 + b^2}} \leq \sqrt{\mathbf{E}_\mu \left(\frac{1}{\left(\frac{\mathbf{L}(h) - a}{b} \right)^2 + 1} \right)}. \tag{53}$$

Using the inequality $1/(x+1) \leq 1-x/2$ for all $0 \leq x \leq 1$ ((36) in Lemma 5), this suggests a natural event to use as a condition. More precisely, writing $\mathbf{E} := \mathbf{I}\{|L(h) - a| \leq b\}$ for readability, note that we have

$$\begin{aligned} \frac{1}{\left(\frac{L(h)-a}{b}\right)^2 + 1} &= \frac{1 - \mathbf{E}}{\left(\frac{L(h)-a}{b}\right)^2 + 1} + \frac{\mathbf{E}}{\left(\frac{L(h)-a}{b}\right)^2 + 1} \\ &\leq \frac{1 - \mathbf{E}}{\left(\frac{L(h)-a}{b}\right)^2 + 1} + \mathbf{E} \left(1 - \frac{1}{2} \left(\frac{L(h)-a}{b}\right)^2\right) \\ &= \underbrace{\left(\frac{1 - \mathbf{E}}{\left(\frac{L(h)-a}{b}\right)^2 + 1} + \mathbf{E}\right)}_{\leq 1} - \frac{\mathbf{E}}{2} \left(\frac{L(h)-a}{b}\right)^2. \end{aligned}$$

Taking expectation and utilizing (53), whenever (51) holds, we have

$$1 - \beta/\lambda = \mathbf{E}_\mu \left(\frac{b}{\sqrt{(L(h)-a)^2 + b^2}} \right) \leq \sqrt{1 - \mathbf{E}_\mu \frac{\mathbf{E}}{2} \left(\frac{L(h)-a}{b}\right)^2}. \quad (54)$$

With this established, note that via helper inequality (37), for any $\beta \leq \lambda$ we have

$$(1 - \beta/\lambda)^2 \geq 1 - 2\beta/\lambda$$

and thus in light of (54), we may conclude that

$$1 - 2\beta/\lambda \leq 1 - \mathbf{E}_\mu \frac{\mathbf{E}}{2} \left(\frac{L(h)-a}{b}\right)^2$$

which implies

$$\frac{\lambda}{4\beta} \mathbf{E}_\mu \mathbf{E}(h; a, b)(L(h) - a)^2 \leq b^2$$

noting that we have written $\mathbf{E}(h; a, b)$ to emphasize the dependence on h , a , and b . Once again since the first-order condition (51) is necessary for optimality, we may conclude that

$$\frac{\lambda}{4\beta} \mathbf{E}_\mu \mathbf{E}(h; a, b_\mu(h, a))(L(h) - a)^2 \leq b_\mu^2(h, a) \quad (55)$$

which is the remaining desired inequality. \square

Proof of the limit (14). Recall from the proof of Proposition 2 the first-order optimality condition (51), which is satisfied by any solution $b_\mu(h, a)$ given by (13), i.e., we have

$$\mathbf{E}_\mu \left(\frac{b_\mu(h, a)}{\sqrt{(L(h)-a)^2 + (b_\mu(h, a))^2}} \right) = 1 - \beta/\lambda \quad (56)$$

for any $0 < \beta \leq \lambda$. Defining $g(\beta) := 1 - \beta/\lambda$ and taking any $0 < \beta_2 < \beta_1 \leq \lambda$, clearly we have $g(\beta_1) < g(\beta_2)$ and thus using the equality (56), we must have that $b_\mu(h, a; \beta_2) \geq b_\mu(h, a; \beta_1)$, otherwise it would result in a contradiction of (56). Using this monotonicity, clearly

$$\mathbf{E}(h; a, b_\mu(h, a; \beta_1)) \leq \mathbf{E}(h; a, b_\mu(h, a; \beta_2))$$

and thus

$$\mathbf{E}_\mu \mathbf{E}(h; a, b_\mu(h, a; \beta_1))(L(h) - a)^2 \leq \mathbf{E}(h; a, b_\mu(h, a; \beta_2))(L(h) - a)^2.$$

Applying this to the lower bound in Proposition 2, we have

$$\liminf_{\beta \rightarrow 0^+} b_\mu^2(h, a) \geq \lim_{\beta \rightarrow 0^+} \frac{\lambda}{4\beta} \mathbf{E}_\mu \mathbf{I}(h; a)(L(h) - a)^2 = \infty$$

as desired. \square

Proof of the limit (15). Note that we can easily bound the random variable of interest as

$$0 \leq b\rho\left(\frac{\mathbf{L}(h) - a}{b}\right) = \sqrt{(\mathbf{L}(h) - a)^2 + b^2} - b \leq |\mathbf{L}(h) - a| \quad (57)$$

for any choice of $0 < b < \infty$. Some straightforward calculus shows that

$$\lim_{b \rightarrow \infty} b\rho\left(\frac{\mathbf{L}(h) - a}{b}\right) = 0$$

in a pointwise sense. Since the upper bound in (57) is μ -integrable by assumption, a simple application of dominated convergence (Ash and Doléans-Dade, 2000, Thm. 1.6.9) yields

$$\lim_{b \rightarrow \infty} b \mathbf{E}_\mu \rho\left(\frac{\mathbf{L}(h) - a}{b}\right) = \mathbf{E}_\mu \left[\lim_{b \rightarrow \infty} b\rho\left(\frac{\mathbf{L}(h) - a}{b}\right) \right] = 0$$

as desired. \square

Proof of Proposition 3. From condition (20), since any solution must also be a stationary point (Nesterov, 2004, Thm. 1.2.1), we know that $\mathbf{A}_n := \mathbf{A}_n(h, b)$ must satisfy the first-order condition

$$\frac{\lambda}{n} \sum_{i=1}^n \rho'\left(\frac{\mathbf{L}_i(h) - \mathbf{A}_n}{b}\right) = \alpha$$

which is equivalent to

$$\frac{b}{n} \sum_{i=1}^n \rho'\left(\frac{\mathbf{L}_i(h) - \mathbf{A}_n}{b}\right) = \frac{\alpha}{\lambda} b. \quad (58)$$

Next we make use of the argument developed by Catoni (2012, §2). First note that fixing any $a \in \mathbb{R}$ and $b > 0$, we have

$$\begin{aligned} \mathbf{E} \exp\left(\sum_{i=1}^n \rho'\left(\frac{\mathbf{L}_i(h) - a}{b}\right)\right) &= \mathbf{E} \left[\prod_{i=1}^n \exp\left(\rho'\left(\frac{\mathbf{L}_i(h) - a}{b}\right)\right) \right] \\ &= \prod_{i=1}^n \mathbf{E}_i \exp\left(\rho'\left(\frac{\mathbf{L}_i(h) - a}{b}\right)\right) \\ &\leq \prod_{i=1}^n \mathbf{E}_i \left(1 + \frac{\mathbf{L}_i(h) - a}{b} + \frac{\gamma}{b^2} (\mathbf{L}_i(h) - a)^2 \right) \\ &= \left(1 + \frac{\mathbf{E}_\mu \mathbf{L}(h) - a}{b} + \frac{\gamma}{b^2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2 \right)^n \\ &\leq \exp\left(\frac{n}{b} (\mathbf{E}_\mu \mathbf{L}(h) - a) + \frac{n\gamma}{b^2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2\right). \end{aligned} \quad (59)$$

The second equality above follows from the independence of the training data, and the first inequality uses the upper bound in (4), which is satisfied by ρ given in (6) with $\gamma = 1$, though we leave γ as is to illustrate how more general results are obtained. The third equality just uses the fact that the training data is an IID sample from μ , and the final inequality culminating in (59) just uses the bound $1 + x \leq \exp(x)$. Using Markov's inequality and taking $0 < \delta < 1$, it is straightforward to show that (59) implies a $1 - \delta$ event (over the draw of $\mathbf{Z}_1, \dots, \mathbf{Z}_n$) in which we have

$$\sum_{i=1}^n \rho'\left(\frac{\mathbf{L}_i(h) - a}{b}\right) \leq \frac{n}{b} (\mathbf{E}_\mu \mathbf{L}(h) - a) + \frac{n\gamma}{b^2} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2 + \log(1/\delta).$$

Multiplying both sides by b/n , on the same “good” event, we have

$$\begin{aligned} \frac{b}{n} \sum_{i=1}^n \rho'\left(\frac{\mathbf{L}_i(h) - a}{b}\right) &\leq \mathbf{E}_\mu \mathbf{L}(h) - a + \frac{\gamma}{b} \mathbf{E}_\mu (\mathbf{L}(h) - a)^2 + \frac{b \log(1/\delta)}{n} \\ &= \mathbf{E}_\mu \mathbf{L}(h) - a + \frac{\gamma}{b} (\mathbf{V}_\mu \mathbf{L}(h) + (\mathbf{E}_\mu \mathbf{L}(h) - a)^2) + \frac{b \log(1/\delta)}{n} \end{aligned} \quad (60)$$

where (60) follows from expanding the quadratic term and doing some algebra. With the equality (58) in mind, subtracting a constant from both sides of (60), note that we equivalently have

$$\frac{b}{n} \sum_{i=1}^n \rho' \left(\frac{\mathbf{L}_i(h) - a}{b} \right) - \frac{\alpha}{\lambda} b \leq p(a) \quad (61)$$

where we have defined

$$p(a) := \mathbf{E}_\mu \mathbf{L}(h) - a + \frac{\gamma}{b} (\mathbf{V}_\mu \mathbf{L}(h) + (\mathbf{E}_\mu \mathbf{L}(h) - a)^2) + \frac{b \log(1/\delta)}{n} - \frac{\alpha}{\lambda} b \quad (62)$$

for readability. Note that $p(\cdot)$ in (62) is a polynomial of degree 2, and can be written as

$$p(a) = ua^2 + va + w \quad (63)$$

with coefficients defined as

$$\begin{aligned} u &:= \frac{\gamma}{b} \\ v &:= (-1) \left(1 + \frac{2\gamma \mathbf{E}_\mu \mathbf{L}(h)}{b} \right) \\ w &:= \mathbf{E}_\mu \mathbf{L}(h) + \frac{\gamma}{b} \mathbf{E}_\mu |\mathbf{L}(h)|^2 + \frac{b \log(1/\delta)}{n} - \frac{\alpha}{\lambda} b. \end{aligned}$$

This polynomial has real roots whenever $v^2 - 4uw \geq 0$, and some algebra shows that this is equivalent to

$$0 \leq D \leq 1, \text{ where } D := 4 \left(\left(\frac{\gamma}{b} \right)^2 \mathbf{V}_\mu \mathbf{L}(h) + \frac{\gamma \log(1/\delta)}{n} - \frac{\gamma \alpha}{\lambda} \right). \quad (64)$$

Assuming this holds, denoting by a_+ the smallest root of $p(\cdot)$, i.e., the smallest of satisfying $p(a_+) = 0$, the critical fact of interest to us is that $\mathbf{A}_n \leq a_+$ on the good event of (61). This is valid due to two facts: first, the left-hand side of (61) is a decreasing function of a ; second, due to (58), we know that \mathbf{A}_n is a root of the left-hand side of (61). With this key fact in hand, using the quadratic formula we have

$$\begin{aligned} \mathbf{A}_n &\leq a_+ \\ &= \mathbf{E}_\mu \mathbf{L}(h) + \frac{b}{2\gamma} \left(1 - \sqrt{1 - D} \right) \\ &= \mathbf{E}_\mu \mathbf{L}(h) + \frac{b}{2\gamma} \frac{(1 - \sqrt{1 - D})(1 + \sqrt{1 - D})}{(1 + \sqrt{1 - D})} \\ &= \mathbf{E}_\mu \mathbf{L}(h) + \frac{b}{2\gamma} \frac{D}{(1 + \sqrt{1 - D})} \\ &\leq \mathbf{E}_\mu \mathbf{L}(h) + \frac{b}{2\gamma} D. \end{aligned}$$

Taking the two ends of this inequality chain together and expanding D , we have

$$\mathbf{A}_n \leq \mathbf{E}_\mu \mathbf{L}(h) - 2(\alpha/\lambda)b + 2 \left(\frac{\gamma}{b} \mathbf{V}_\mu \mathbf{L}(h) + \frac{b \log(1/\delta)}{n} \right) \quad (65)$$

with probability no less than $1 - \delta$, assuming that n , b , and α are such that $0 \leq D \leq 1$ holds. This gives us the desired upper bound.

To obtain a lower bound, a perfectly analogous argument can be applied. First, using the *lower* bound in (4) and the fact that $\rho'(-x) = -\rho'(x)$, we know that

$$\rho' \left(\frac{a - \mathbf{L}_i(h)}{b} \right) \leq \log \left(1 + \frac{a - \mathbf{L}_i(h)}{b} + \frac{\gamma}{b^2} (a - \mathbf{L}_i(h))^2 \right) \quad (66)$$

for any $a \in \mathbb{R}$, $b > 0$, and $i \in [n]$. Plugging this inequality (66) into an argument analogous to the chain of inequalities that led to (60) earlier, it is clear that again on an event of probability no less than $1 - \delta$, we have

$$\frac{b}{n} \sum_{i=1}^n \rho' \left(\frac{a - \mathbf{L}_i(h)}{b} \right) \leq a - \mathbf{E}_\mu \mathbf{L}(h) + \frac{\gamma}{b} (\mathbf{V}_\mu \mathbf{L}(h) + (\mathbf{E}_\mu \mathbf{L}(h) - a)^2) + \frac{b \log(1/\delta)}{n}. \quad (67)$$

Once again the upper bound we can bound this using a polynomial of degree 2, namely

$$\frac{b}{n} \sum_{i=1}^n \rho' \left(\frac{a - \mathbf{L}_i(h)}{b} \right) + \frac{\alpha}{\lambda} b \leq q(a) \quad (68)$$

where we have defined

$$q(a) := a - \mathbf{E}_\mu \mathbf{L}(h) + \frac{\gamma}{b} (\mathbf{V}_\mu \mathbf{L}(h) + (\mathbf{E}_\mu \mathbf{L}(h) - a)^2) + \frac{b \log(1/\delta)}{n} + \frac{\alpha}{\lambda} b. \quad (69)$$

Now, since \mathbf{A}_n is a root of the left-hand side of (68) viewed as a function of a , and this function is monotonically increasing, it is evident that denoting the largest root of $q(\cdot)$ (when it exists) by a_- , we have $\mathbf{A}_n \geq a_-$, a lower bound in contrast to the $\mathbf{A}_n \leq a_+$ upper bound used earlier. For completeness, we write this polynomial as

$$q(a) = u'a^2 + v'a + w' \quad (70)$$

with coefficients

$$\begin{aligned} u' &:= \frac{\gamma}{b} \\ v' &:= \left(1 - \frac{2\gamma \mathbf{E}_\mu \mathbf{L}(h)}{b} \right) \\ w' &:= (-1) \mathbf{E}_\mu \mathbf{L}(h) + \frac{\gamma}{b} \mathbf{E}_\mu |\mathbf{L}(h)|^2 + \frac{b \log(1/\delta)}{n} + \frac{\alpha}{\lambda} b. \end{aligned}$$

We have two real roots whenever

$$1 \geq D' := 4 \left(\left(\frac{\gamma}{b} \right)^2 \mathbf{V}_\mu \mathbf{L}(h) + \frac{\gamma \log(1/\delta)}{n} + \frac{\gamma \alpha}{\lambda} \right) \quad (71)$$

holds, and thus we obtain a high probability lower bound on \mathbf{A}_n as follows:

$$\begin{aligned} \mathbf{A}_n &\geq a_- \\ &= \mathbf{E}_\mu \mathbf{L}(h) - \frac{b}{2\gamma} \left(1 - \sqrt{1 - D'} \right) \\ &\geq \mathbf{E}_\mu \mathbf{L}(h) - \frac{b}{2\gamma} D'. \end{aligned}$$

Expanding D' gives us the lower bound

$$\mathbf{A}_n \geq \mathbf{E}_\mu \mathbf{L}(h) - 2(\alpha/\lambda)b - 2 \left(\frac{\gamma}{b} \mathbf{V}_\mu \mathbf{L}(h) + \frac{b \log(1/\delta)}{n} \right) \quad (72)$$

with probability no less than $1 - \delta$, as desired.

Let us conclude this proof by organizing the technical assumptions. First of all, for the two quadratics used in the preceding bounds, we require both (64) and (71) to hold. It is straightforward to verify that having these conditions both hold is equivalent to the following:

$$\frac{4\gamma\alpha}{\lambda} \leq 4 \left(\left(\frac{\gamma}{b} \right)^2 \mathbf{V}_\mu \mathbf{L}(h) + \frac{\gamma \log(1/\delta)}{n} \right) \leq 1 - \frac{4\gamma\alpha}{\lambda}. \quad (73)$$

As such, whenever α , δ , and b are such that (73) holds, using a union bound, it follows that with probability no less than $1 - 2\delta$, we have a bound on

$$|\mathbf{A}_n - (\mathbf{E}_\mu \mathbf{L}(h) - 2(\alpha/\lambda)b)| \leq 2 \left(\frac{\gamma}{b} \mathbf{V}_\mu \mathbf{L}(h) + \frac{b \log(1/\delta)}{n} \right)$$

as desired. The proposition statement takes a cleaner form since we have $\gamma = 1$. \square

Proof of Proposition 4. The lack of convexity follows from the fact that the composition of two convex functions need not be convex when the outermost function is *non-monotonic* (see for example [Boyd and Vandenberghe \(2004, Ch. 3\)](#)), and the lack of smoothness follows *a fortiori* from Lemma 7. \square