Risk regularization through bidirectional dispersion

Matthew J. Holland*
Osaka University

Abstract

Many alternative notions of “risk” (e.g., CVaR, entropic risk, DRO risk) have been proposed and studied, but these risks are all at least as sensitive as the mean to loss tails on the upside, and tend to ignore deviations on the downside. In this work, we study a complementary new risk class that penalizes loss deviations in a bidirectional manner, while having more flexibility in terms of tail sensitivity than is offered by classical mean-variance, without sacrificing computational or analytical tractability.

Contents

1 Introduction 2

2 An initial look at regularized risks 4
    2.1 Dispersion-based risk families 4
    2.2 A flexible class of symmetric dispersion functions 5
    2.3 Connections with traditional location parameters 6
    2.4 Alternative risk classes 7

3 Basic theory 9

4 Learning theory 12
    4.1 Measuring performance 13
    4.2 Unbiased stochastic feedback 13
    4.3 Smoothness of the objective 14
    4.4 A concrete learning algorithm 16
    4.5 Stationarity guarantees 16
    4.6 Moving beyond the standard approach 17

5 Empirical analysis 18
    5.1 Simulation-based risk class analysis 19
        5.1.1 Experimental setup 19
        5.1.2 Results and discussion 19
    5.2 Loss distributions at initialization 21
    5.3 Outliers and risk design 21
        5.3.1 Experimental setup 22
        5.3.2 Results and discussion 22
    5.4 Classification error trajectories 23
        5.4.1 Experimental design 23

*Please direct correspondence to matthew-h@ar.sanken.osaka-u.ac.jp.
1 Introduction

What does it mean for a learner to generalize well? Broadly speaking, this is an ambiguous property of learning systems that can be defined, measured, and construed in countless ways. In the context of machine learning, however, the notion of “good off-sample generalization” is almost without exception formalized as minimizing the expected value of a random loss $E_\mu L(h)$, where $h$ is a candidate parameter or decision rule, and $L(h)$ is a random variable on a probability space $(\Omega, \mathcal{F}, \mu)$ [41, 49]. The idea of quantifying the risk of an unexpected outcome (here, a random loss) using the expected value dates back to the Bernoullis and Gabriel Cramer in the early 18th century [5, 23]. In a more modern context, the emphasis on average performance is the “general setting of the learning problem” of Vapnik [51], and plays a central role in the decision-theoretic learning model of Haussler [25]. Use of the expected loss to quantify off-sample generalization has been essential to the development of both the statistical and computational theories of learning [16, 31].

While the expected loss has a long history and remains pervasive, several new lines of work have begun exploring novel feedback mechanisms for learning algorithms, in some cases based on alternative risk functions that generalize or replace the expected loss. The idea of designing objective functions that explicitly penalize both location and deviation is now classical, with the “mean-variance” objective $E_\mu L + E_\mu(L - E_\mu L)^2$ being the canonical example that formed the basis for modern portfolio optimization theory [39]. Insights from statistical learning theory tell us that smaller variance can imply better guarantees for empirical risk minimizers [9, §5], and such insights have been leveraged to show the statistical advantages of variance-based risk regularization [40]. Note that the variance penalizes deviations about the mean in a symmetric, bidirectional fashion; tails on the upside have as much influence as tails on the downside.

On the other hand, many risk functions studied in recent years are asymmetric in their treatment of loss deviations. One of the best-known and most influential examples is conditional value-at-risk (CVaR) [45, 46]. Early work in machine learning using CVaR includes the design of cost-sensitive learning algorithms [30] and formal connections to certain support vector
machine variants [50]. More recent work has considered the design and analysis of stochastic learning algorithms [12, 13, 27], relations to fairness [53], and boosted variants specialized to classification [55]. CVaR is the expected loss conditioned on the event that the loss exceeds a pre-specified quantile. As such, deviations on the downside (below the quantile threshold) are completely ignored. Another alternative called the entropic risk uses exponential smoothing to approximate extreme quantiles of the loss distribution on the upside [19, 34]. This is an established technique in reinforcement learning [20], whose extension and application to general-purpose empirical risk minimization (ERM) has recently been studied in depth under the name tilted ERM (TERM) [35]. Both CVaR and entropic risk are special cases of the optimal certainty equivalent (OCE) risk [7, 8, 34], defined as the sum of a generalized location parameter of the true loss distribution and a term that penalizes deviation of the loss around this location in an asymmetric fashion (see §2.4 for details).

Convex relaxations of the mean-variance risk have also received significant attention. In particular, robustly regularized risks taking the form \( h \mapsto \sup \{ E_{\mu} L(h) : L \in \mathcal{L} \} \), where \( \mathcal{L} \) is a family of random losses reflecting all possible deviations from an underlying data distribution, up to a certain threshold typically given in terms of a Rényi (or Cressie-Read) divergence [17, 18]. This relaxation places the problem within the realm of (stochastic) distributionally robust optimization (DRO) [6], a well-studied area in which clear links between mean-variance and robustly regularized risks are known [22, 42]. As with the OCE risks described previously, the convex relaxation involved in robust regularization yields a risk function that penalizes deviations on the upside, while ignoring the downside.

In this work, our goal is to investigate properties of the loss distribution which are not captured by the preceding risk notions, and to consider the feasibility and utility of using such properties as (noisy) feedback for learning algorithms. More concretely, we consider a class of risks that penalize deviations in a bidirectional manner, while having more flexibility to tail sensitivity than is offered by classical mean-variance (§2–§3). We develop formal learning guarantees under first-order stochastic optimizers (§4), and complement this with a rigorous empirical analysis in which we compare and contrast our risk classes with the alternatives mentioned in the preceding paragraphs (§5). We pay particular attention to the similarities and differences that arise as we modify the sensitivity of each risk class to the tails of the underlying loss distribution. This analysis is not limited to just comparing the risk values computed by each class, but also by studying how loss distributions change under learning algorithms that take each risk as their ultimate objective. Our work is conceptually closest to the work of Li et al. [35, 36], though our theoretical focus is a completely different class of risks, and our novel experimental design for exploring risk class behavior (especially in §5.1) offers a new perspective into risk classes new and old.\(^1\)

**Notation** Here we summarize the conventions we use for our notation in this paper. To start, let us clarify the nature of the random losses we consider. With the context of the underlying probability space \((\Omega, \mathcal{F}, \mu)\), we will write \( L(h) := L(h; \cdot) : \Omega \to \mathbb{R} \) to refer to a random variable (i.e., a \( \mathcal{F} \)-measurable function) on \( \Omega \), though we will only use the form \( L(h) \) in the body of this paper. When we talk about “sampling” losses or a “random draw” of the losses, this amounts to computing a realization \( L(h; \omega) \in \mathbb{R} \). We use standard notation for taking expectation, e.g., \( E_{\mu} L(h) := \int_{\Omega} L(h; \omega) \mu(d\omega) \). These conventions extend to random quantities based on the losses (e.g., the gradient \( L'(h) \) considered in §4). When we write \( E \) without the subscript \( \mu \), this will typically refer to expectation taken over all the random objects available (i.e., the product measure induced by a random sample). Similarly, we will use \( P \) as a general-purpose

\(^1\)We have included a key subset of the TERM risk class in our empirical analysis, denoted “entropic risk.”
probability function, representing both \( \mu \) and product measures; the underlying randomness should be immediate from the context. We will use \( \| \cdot \| \) as a general-purpose notation for all norms that appear in this paper. That is, we do not use different notation to distinguish different norm spaces. The reason for this is that we will never consider two distinct norms on the same set; each norm is associated with a distinct set, and thus as long as it is clear which set a particular element belongs to, there should be no confusion. The only exception to this rule is the special case of \( \mathbb{R} \), in which we write \(| \cdot |\) for the absolute value, as is traditional. For a function \( f : \mathbb{R} \to \mathbb{R} \) in one variable, we use \( f' \) to denote the usual derivative. More general notions (e.g., Gateaux or Fréchet differentials) only make an appearance in §4, and the generality they afford us is not crucial to the main narrative, so the details can be easily skipped over if the reader is unfamiliar with such concepts. All the other undefined notation we use is essentially standard, and can be found in most introductory analysis textbooks.

2 An initial look at regularized risks

In this section, we begin by formulating the notion of dispersion functions, and use this to define two important risk classes (§2.1). We then introduce a flexible class of symmetric dispersion functions (§2.2), discuss links to traditional location parameters (§2.3), and review important existing risk classes with a particular emphasis on the role of dispersion underlying each (§2.4). All formal matters such as existence, finite moments, differentiability, and smoothness are relegated to §3, with technical details in §A.2.

2.1 Dispersion-based risk families

To ground ourselves conceptually, let us refer to \( \text{L}(h) \) as the base loss incurred by \( h \). The exact nature of \( h \) here is left completely abstract; all that matters is the probability distribution of the base loss. While this underlying distribution has many properties that can in principle be used to design risk functions and subsequently generate feedback (i.e., transformed losses), in this work we pay particular attention to dispersion (or “deviation”) properties of the loss.\(^2\)

Let us thus start by defining

\[
\text{D}_f(h; \theta) := \mathbb{E}_\mu f \left( \frac{\text{L}(h) - \theta}{\sigma} \right) \tag{1}
\]

where \( \sigma > 0 \) is a scaling parameter, and \( f : \mathbb{R} \to \mathbb{R}_+ \) is a function taking non-negative real values, controlling how we measure the loss dispersion at different scales (relative to \( \sigma \)) and in each direction (above or below \( \theta \)). Writing \( f_\sigma(x) := f(x/\sigma) \) for readability, we call \( f_\sigma(\text{L}(h) - \theta) \) the (random) dispersion of the base loss, taken with respect to threshold \( \theta \). Thus \( \text{D}_f(h; \theta) \) defined in (1) is simply the expected dispersion. For ease of reference, we shall use the term dispersion function to refer to any function (here, \( f \)) used when defining a dispersion through (1). Since \( f \) is bounded below, we always have a finite infimum

\[
\text{D}_f(h) := \inf \{ \text{D}_f(h; \theta) : \theta \in \mathbb{R} \} \tag{2}
\]

which we call the M-dispersion of the loss incurred by \( h \).\(^3\)

Moving forward, we consider two risk families that are characterized by their dependence on the base loss dispersion. The first class starts with a traditional location parameter (here,

---

\(^2\)We use the term “dispersion” in order to distinguish the concept from the well-known class of general deviation functions studied by Rockafellar et al. [47]. We discuss this point in more detail after Lemma 7 in §3.

\(^3\)Traditionally, an “M-parameter” of a probability distribution is characterized as a solution to a minimization problem [33], but this concept naturally extends to the optimal function values achieved by M-parameters.
the expected loss) and simply uses the M-dispersion as an additive regularizer. We call this an \textit{M-regularized risk} (or \textit{M-risk} for short), defined as

\[ R_f(h) := \mathbb{E}_\mu L(h) + \eta D_f(h) \]  

with weighting parameter \( \eta \geq 0 \). In contrast to this, the second class we consider begins with the expected dispersion, and implicitly defines a location though the form

\[ \tilde{R}_f(h) := \inf \{ D_f(h; \theta) + \tilde{\eta}\theta : \theta \in \mathbb{R} \} \]  

using weighting parameter \( \tilde{\eta} \in \mathbb{R} \), which is allowed to be negative. We refer to any member of this risk family as \textit{threshold risk} (or \textit{T-risk} for short). The definitions in (3) and (4) represent classes of risk functions, parameterized by the choice of \( \sigma, f, \) and \( \eta \) or \( \tilde{\eta} \).

\[ \text{Remark 1 (Two risk classes).} \] Clearly the M-risk and T-risk given in (3)–(4) are closely related, but they differ fundamentally in the role played by their “location” terms. With \( R_f \), the risk is a typical sum of location and dispersion, where the location term and the threshold about which dispersion is measured are not the same in general. On the other hand, \( \tilde{R}_f \) has a location which is determined implicitly by the dispersion function used, so by definition the location term aligns precisely with the dispersion threshold. If \( \tilde{\eta} > 0 \), then one uses a threshold below the M-threshold for computing dispersion, and assigns a larger risk to larger thresholds. On the other hand, if \( \tilde{\eta} < 0 \), then one uses a threshold above the M-threshold for computing dispersion, and uses this threshold to reduce the risk. At first glance, this may not seem intuitive, since losses come with the interpretation that “lower is better,” and in principle a loss distribution with arbitrarily bad location could be made optimal as long as it is sufficiently well-concentrated (i.e., small dispersion). We show in §2.3 that this risk class comes from a very straightforward generalization of the convex program used to characterize loss quantiles, which offers greater control over sensitivity to the distribution tails.

\subsection*{2.2 A flexible class of symmetric dispersion functions}

With preliminary definitions in place, recall from §1 that our main interest in this work is “two-way” dispersion functions which treat dispersion on both the upside and the downside in approximately the same way. We will focus on a particular class of symmetric dispersion functions, here denoted by \( \rho \), that can be plugged in to either (3) or (4), setting \( f = \rho \). With shape parameter \(-\infty \leq \alpha \leq 2\), we consider the following flexible class of functions:

\[ \rho(x; \alpha) := \begin{cases} 
\frac{x^2}{2}, & \text{if } \alpha = 2 \\
\log(1 + \frac{x^2}{2}), & \text{if } \alpha = 0 \\
1 - \exp\left(-\frac{x^2}{2}\right), & \text{if } \alpha = -\infty \\
\frac{\lvert\alpha-2\rvert}{\alpha}\left(1 + \frac{x^2}{\lvert\alpha-2\rvert}\right)^{\alpha/2} - 1, & \text{otherwise}
\end{cases} \]  

Whenever we say that a dispersion function \( \rho \) belongs to the Barron class, we mean that for some \(-\infty \leq \alpha \leq 2\), we have \( \rho(\cdot) = \rho(\cdot; \alpha) \) following the definition in (5). At a high level, this function is approximately quadratic near zero for any choice of shape parameter \( \alpha \), but its growth as one deviates far from zero can be substantially modified via \( \alpha \) (see Figure 1).

\[ \text{For readability, we will often suppress the dependence on } \sigma, \eta, \text{ and } \tilde{\eta} \text{ in our notation.} \]

\[ \text{The reason for this naming is that Barron [4] recently studied this class in the context of designing loss functions for computer vision applications. We remark that this differs considerably from our usage in computing the dispersion of random losses, where the loss function underlying the base loss is left completely arbitrary.} \]
Figure 1: Here we consider the function \( x \mapsto \rho(x/\sigma; \alpha) \), with \( \rho \) from the Barron class \((5)\), for a variety of shape parameter values \( \alpha \), with scale parameter \( \sigma = 0.5 \) fixed. We plot graphs for this function (left), its first derivative (center), and its second derivative (right), each computed over \( \pm 5\sigma \).

Clearly, the function \( x \mapsto \rho(x; \alpha) \) is symmetric about zero, since \( \rho(x; \alpha) = \rho(-x; \alpha) \) for any \( x \in \mathbb{R} \), and thus it can be used to measure dispersion in a bidirectional manner. For any valid choice of \( \alpha \), we have \( \rho(0; \alpha) = 0 \), and \( \rho(x; \alpha) > 0 \) for all \( x \neq 0 \). It is twice continuously differentiable on \( \mathbb{R} \) for any valid \( \alpha \) (see §B.3 for exact expressions). All the limits in \( \alpha \) behave as we would expect: \( \rho(x; \alpha) \to \rho(x; c) \) as \( \alpha \to c \) for \( c \in \{-\infty, 0, 2\} \) (see §A.1.1 for details).

For \( \alpha \geq 0 \), the dispersion function is unbounded, with growth ranging from logarithmic to quadratic depending on the choice of \( \alpha \). For \( \alpha < 0 \), the dispersion function is bounded. By controlling boundedness and growth rates, we can modify the sensitivity of the expected dispersion to the tails of the base loss distribution.

2.3 Connections with traditional location parameters

To further develop our intuition for the dispersion and risk classes described in the preceding sub-sections, let us consider a concrete re-scaling strategy. Let \( \rho \) denote a dispersion function from the Barron class \((5)\). As mentioned earlier, one important property of this class is that around zero, the function is approximately quadratic. Intuitively, taking \( \sigma \) relatively large will imply that the dispersion is measured in an approximately quadratic fashion. In fact, with proper re-scaling, we can ensure that the dispersion is indeed quadratic in the limit of \( \sigma \to \infty \).

More precisely, it is straightforward to verify that we have

\[
\lim_{\sigma \to \infty} \sigma^2 \rho(x; \alpha) = x^2
\]

for any choice of \(-\infty \leq \alpha \leq 2\) and \( x \in \mathbb{R} \). What about the other direction, in which we take \( \sigma \to 0_+ \)? First note that if \(-\infty < \alpha < 2\) and \( \alpha \neq 0 \), for any \( c \geq 0 \) multiplying \( c\sigma^\alpha \) by the dispersion gives us

\[
c\sigma^\alpha \rho(x; \alpha) = \frac{\left|\alpha - 2\right|}{\alpha} \left(\sigma^2 + \frac{x^2}{\left|\alpha - 2\right|}\right)^{\alpha/2} - \sigma^\alpha c
\]

and thus whenever \( \alpha > 0 \), it immediately follows that \( c\sigma^\alpha \rho(x; \alpha) \to \alpha^{-1}\left|\alpha - 2\right|^{-\alpha/2}c|x|^\alpha \) as \( \sigma \to 0_+ \). As such, for any \( 0 < \alpha < 2 \), it follows that

\[
\lim_{\sigma \to 0_+} \eta_\sigma \rho(x; \alpha) = |x|^{\alpha}, \text{ where } \eta_\sigma := \alpha\left|\alpha - 2\right|^{-\alpha/2}^{-1}\sigma^\alpha.
\]

Considering (6) and (7) together, we see that under \( \rho \) from the Barron class, with appropriate settings of \( \eta \) (in \( R_\rho \)) and \( \tilde{\eta} \) (in \( \tilde{R}_\rho \)) depending on \( \sigma > 0 \) and \( \alpha > 0 \), it is possible to ensure that
the re-scaled dispersion interpolates between sub-quadratic polynomial behavior ($\approx |x|^{\alpha}$) and quadratic behavior ($\approx x^2$).\footnote{The requirement of $\alpha > 0$ is needed for the $\sigma \to 0_+$ direction, since if $\alpha < 0$, then $\sigma^\alpha \to \infty$ as $\sigma \to 0_+$.}

To put these basic facts in a slightly broader context, taking $\beta \in (0, 1)$ let us define the \textit{$\beta$-quantile} of the loss distribution as

$$Q_\beta(h) := \inf\{x \in \mathbb{R} : P\{L(h) \leq x\} \geq \beta\}. \tag{8}$$

An important property of quantiles is that they can be characterized as solutions to a convex optimization problem. More precisely, whenever $E_\mu[L(h)] < \infty$, it can be shown that

$$Q_\beta(h) \in \arg \min_{\theta \in \mathbb{R}} E_\mu[L(h) - \theta] + (1 - 2\beta)\theta \tag{9}$$

for any choice of $0 < \beta < 1$ [33]. As such, the threshold risk $\tilde{R}_\rho$ under the Barron dispersion class can be seen as a straightforward generalization of the convex program characterizing quantiles. Indeed, setting $\alpha = 1$ and letting $\tilde{\eta} = (1 - 2\beta)/\sigma$, in light of (7) it is clear that taking $\sigma$ sufficiently small, the optimal threshold characterizing $\tilde{R}_\rho(h)$ can be made to approach any $\beta$-quantile $Q_\beta(h)$.

On the other end of the spectrum, expected value can be similarly characterized as

$$E_\mu L(h) \in \arg \min_{\theta \in \mathbb{R}} E_\mu (L(h) - \theta)^2 \tag{10}$$

whenever $E_\mu[L(h)]^2 < \infty$. We can obviously recover the expected value exactly using $R_\rho$ by setting $\eta = 0$, or use it as a threshold by setting $\eta > 0$ and taking $\alpha = 2$ or simply taking $\sigma$ large enough and re-scaling as described in (6). As for $\tilde{R}_\rho$, considering (6) and (7), through proper re-scaling it is clear that we have the flexibility to shift between tail-insensitive quantiles and values near the mean that are more tail-sensitive.

### 2.4 Alternative risk classes

Recalling the risk and dispersion functions we formulated in §2.1–§2.2, the foundation for our subsequent analysis will be the M-risk $R_\rho$ and T-risk $\tilde{R}_\rho$, taking $\rho$ from the Barron class (5). In order to facilitate a precise comparison between our risks of interest and important existing risk classes, here we provide a succinct review of risks that appear prominently in the statistical machine learning literature.

To begin, we start with a class of risks whose definition is very similar to that of the T-risk described in (4). Let $\phi : \mathbb{R} \to \mathbb{R}$ be a non-decreasing, closed, convex function which satisfies both $\phi(0) = 0$ and $1 \in \partial \phi(0)$. Given $\phi$, the \textit{OCE risk} is defined as

$$R_{\text{OCE}}(h) := \inf\{\theta + E_\mu \phi(L(h) - \theta) : \theta \in \mathbb{R}\}. \tag{11}$$

This risk was originally developed in the context of expected utility theory for decision-making under uncertainty [7, 8], but can be readily adapted to machine learning settings [34]. Note that if we denote $\varphi(x) := \phi(x) - x$, the OCE risk can clearly be re-written as

$$R_{\text{OCE}}(h) = E_\mu L(h) + \inf\{E_\mu \varphi(L(h) - \theta) : \theta \in \mathbb{R}\} = E_\mu L(h) + D_\varphi(h). \tag{12}$$

That is, if $\theta$ is such that $D_\rho(h; \theta) + \tilde{\eta}\theta = \tilde{R}_\rho(h)$, then $\theta \approx Q_\beta(h)$.
As such, any OCE risk can always be interpreted as both an M-risk \((3)\) and T-risk \((4)\), which are characterized by the “dispersion” functions \(\varphi\) and \(\phi\) respectively. There are three special cases of OCE risk that are of particular importance, which we already briefly mentioned in §1. First, setting \(\phi(x) = x\) we recover the traditional mean \(\text{R}_{\text{OCE}}(h) = E_{\mu}L(h)\). Second, setting \(\phi(x) = (e^{\gamma x} - 1)/\gamma\) for any \(\gamma > 0\), we obtain the \textit{entropic risk}, given as

\[
\text{R}_{\text{OCE}}(h) = \inf \left\{ \theta + \frac{1}{\gamma} \left( E_{\mu} e^{\lambda(h) - \theta} - 1 \right) \right\} = \frac{1}{\gamma} \log \left( E_{\mu} e^{\gamma L(h)} \right). \tag{13}
\]

It should be noted that the equality \((13)\) in fact holds for any \(\gamma \neq 0\), but when \(\gamma < 0\) this is not an OCE risk.\(^8\) Finally, taking any \(0 \leq \beta < 1\) and setting \(\phi(x) = \max\{0, x\}/(1 - \beta)\) we obtain the \textit{conditional value-at-risk} (CVaR) \([45]\), namely we have

\[
\text{R}_{\text{OCE}}(h) = \inf \left\{ \theta + \frac{1}{1 - \beta} E_{\mu}(L(h) - \theta)_+ \right\} = E_{\mu}\left[ L(h) | L(h) \geq Q_{\beta}(h) \right] \tag{14}
\]

where the \(\beta\)-quantile \(Q_{\beta}\) is as defined in \((8)\).

Another important class of risk functions is that of \textit{robustly regularized} risks which are designed to ensure that risk minimizers are robust to a certain degree of divergence in the underlying data model. Making this concrete, it is typical to assume the random losses are the outputs of a loss function \(\ell\) depending on the candidate \(h\) and some random data \(Z\), i.e., \(L(h) = \ell(h; Z)\), with \(Z \sim \mu\) as our reference model. To measure divergence from this reference model, it is convenient to use the Cressie-Read family of divergence functions \([17, 54]\), namely for any \(c > 1\) and assuming \(\nu \ll \mu\) (absolutely continuity) holds, functions of the form

\[
\text{Div}_c(\nu; \mu) := E_{\mu} f_c \left( \frac{d\nu}{d\mu} \right), \text{ where } f_c(x) := \frac{x^c - cx + c - 1}{c(c - 1)} \tag{15}
\]

and \(d\nu/d\mu\) is the Radon-Nikodym density of \(\nu\) with respect to \(\mu\).\(^9\) The resulting robustly regularized risk, called the \textit{DRO risk}, is defined as

\[
\text{R}_{\text{DRO}}(h) := \sup \{ E\, L(h) : L \in \mathcal{L} \} \tag{16}
\]

where the constrained set of random losses \(\mathcal{L}\), determined by \(c > 1\) and \(a > 0\), is defined as

\[
\mathcal{L} := \{ L(h) = \ell(h; Z) : Z \sim \nu \text{ and } \text{Div}_c(\nu; \mu) \leq a \}. \tag{17}
\]

For this particular family of divergences, the risk can be characterized as the optimal value of a simple optimization problem \([17]\), namely we have that

\[
\text{R}_{\text{DRO}}(h) = \inf \left\{ \theta + (1 + c(c - 1)a)^{1/c} \left( E_{\mu} \left( L(h) - \theta \right)_+ \right)^{1/c_a} : \theta \in \mathbb{R} \right\} \tag{18}
\]

where \(c_a := c/(c - 1)\). While strictly speaking this is not an OCE risk, note that if we set \(\phi(x) = (1 + c(c - 1)a)^{c_a/c}(x)^{c_a}_+\), then the DRO risk can be written as

\[
\text{R}_{\text{DRO}}(h) = \inf \{ \theta + [E_{\mu} \phi(L(h) - \theta)]^{1/c_a} : \theta \in \mathbb{R} \},
\]

\(^8\)To see this, just note that \(\phi''(x) = \gamma \exp(\gamma x)\), and thus \(\gamma < 0\) implies \(\phi\) is concave.

\(^9\)For background on absolute continuity and density functions, see Ash and Doléans-Dade [3, §2.2].
giving us an expression of this risk as the sum of a threshold and an asymmetric dispersion. When we set \( c = 2 \), this yields the well-known special case of \( \chi^2 \)-DRO risk [24, 54]. In addition to the one-directional nature of the dispersion term in these risks, all of these risks are at least as sensitive to loss tails (on the upside) as the classical expected loss \( E_{\mu} L(h) \) is; this holds for CVaR (with \( \beta > 0 \)), entropic risk (with \( \gamma > 0 \)), and even robust variants of DRO risk [54].

**Other notions of risk** While our empirical analysis in §5 will be restricted to the risk functions described in the preceding paragraphs, there still exist numerous alternatives, which we briefly describe here. One natural alternative is to use (population) M-estimators of location as risk, i.e., replacing the “inf” on the right-hand side of (2) with “arg min.” This is conceptually perfectly natural, and amounts to using a location induced by a dispersion function, rather than a sum of location and dispersion as we have done in (3) and (4). From a purely statistical perspective, this can be used to reduce the impact of outliers and construct a robust alternative to naive ERM [11], but computational difficulties make it impractical. Another natural choice is a weighted sum of base loss quantiles, also known as (population) L-statistics [29] or spectral risks [1]. This class is extremely broad, and CVaR is the special case that we have selected for comparison, but countless other weighting strategies can be conceived of.

### 3 Basic theory

In this section, we look at core properties of the risks \( R_{\rho} \) and \( \tilde{R}_{\rho} \) under the Barron class of dispersion functions. Before dealing with random losses, let us first summarize basic convexity and smoothness properties of these dispersion functions, taken at arbitrary scale.

**Lemma 2** (Dispersion function convexity and smoothness). Consider \( \rho_\sigma(x) := \rho(x/\sigma; \alpha) \) with \( \rho(\cdot; \alpha) \) from the Barron class (5), with parameter \(-\infty \leq \alpha \leq 2\). The following properties hold for any choice of \( \sigma > 0 \).

- **Case of \( \alpha = 2 \):**
  \( \rho_\sigma \) is convex and \( 1/\sigma^2 \)-smooth on \( \mathbb{R} \).

- **Case of \( \alpha = 0 \):**
  \( \rho_\sigma \) is convex on \( [-\sqrt{2}\sigma, \sqrt{2}\sigma] \), and is \( 1/(\sqrt{2}\sigma) \)-Lipschitz and \( 1/\sigma^2 \)-smooth on \( \mathbb{R} \).

- **Case of \( \alpha = -\infty \):**
  \( \rho_\sigma \) is convex on \( [-\sigma, \sigma] \), and is \((1/\sigma)\exp(-1/2)\)-Lipschitz and \( 1/\sigma^2 \)-smooth on \( \mathbb{R} \).

- **Otherwise:**
  \( \rho_\sigma \) is \( 1/\sigma^2 \)-smooth on \( \mathbb{R} \). When \( \alpha \geq 1 \), \( \rho_\sigma \) is convex on \( \mathbb{R} \). When \( \alpha = 1 \), \( \rho_\sigma \) is also \( 1/\sigma \)-Lipschitz on \( \mathbb{R} \). Else, when \( \alpha < 1 \), we have that \( \rho_\sigma \) is convex between \( \pm \sigma \sqrt{\alpha - 2}/(1 - \alpha) \), and is \((1/\sigma)(\sqrt{(1 - \alpha)}/|\alpha - 2|)^{1-\alpha} \)-Lipschitz on \( \mathbb{R} \).

Furthermore, all these coefficients are tight (see also Figure 2).

---

10 It goes without saying that one can express quantiles below the mean by “inverting” CVaR or TERM (see for example [34, §2.2] or TERM with negative tilt parameter [35]). That said, one must decide whether or not to aim above or below the mean before looking at the data. Using bidirectional dispersion effectively automates this decision.

11 In principle, a similar idea can be applied to the population variants of all sorts of robust mean estimators. See Lugosi and Mendelson [38] for a modern survey of robust mean estimators.

12 Basic statistical learning theory under a large sub-class of spectral risks has been developed recently [28, 32].
Here we show zoomed-in versions of the two right-most plots in Figure 1, now with dashed horizontal lines denoting the Lipschitz (left) and smoothness (right) coefficients stated in Lemma 2 (both positive and negative values). The Lipschitz coefficients depend on $\alpha$ and $\sigma$, and the colors in the left plot reflect the $\alpha$ value. The smoothness coefficient $1/\sigma^2$ is independent of $\alpha$, and is drawn in light gray. Numerically, we can see that the coefficients are tight, though tighter lower bounds for the second derivatives are possible.

Both $R_\rho$ and $\tilde{R}_\rho$ are defined using the expected dispersion $D_\rho(h; \theta)$, and thus in order to ensure the risks are well-defined or finite, such properties for the expected dispersion need to be established. The following lemma gives conditions characterizing when the random dispersion is $\mu$-integrable.

**Lemma 3.** Let $\rho$ be as in Lemma 2. Then to ensure $D_\rho(h; \theta) < \infty$ holds for all $\theta \in \mathbb{R}$, each of the following conditions is sufficient.

- For $0 < \alpha \leq 2$, let $E_\mu|L(h)|^\alpha < \infty$.
- For $\alpha = 0$, let $E_\mu|L(h)|^c < \infty$ for some $c > 0$.
- For $-\infty \leq \alpha < 0$, let $L(h)$ be $\mathcal{F}$-measurable.

Furthermore, for the cases where $\alpha \neq 0$, the above conditions are also necessary.

**Remark 4.** The preceding Lemma 3 tells us when the expected dispersion is finite, and thus it is straightforward to extend this property to the risks $R_\rho$ and $\tilde{R}_\rho$. For the M-risk $R_\rho$, clearly we require $|E_\mu L(h)| < \infty$, and thus assuming $E_\mu|L(h)|^c < \infty$ with $c \geq 1$ is sufficient to ensure $|R_\rho(h)| < \infty$ for all $\alpha \leq c$. The T-risk $\tilde{R}_\rho$ is only slightly more complicated. First note that for any $-\infty < \alpha < 0$, direct inspection shows us that $\sup\{\rho(x; \alpha) : x \in \mathbb{R}\} = (\alpha - 2)/\alpha < \infty$, and thus trivially $\tilde{R}_\rho(h) = -\infty$ in this case. More generally, even though $\sup\{\rho(x; \alpha) : x \in \mathbb{R}\} = \infty$ for all $\alpha \geq 0$, the dispersion function grows too slowly when $\alpha < 1$, and thus we can only guarantee $|\tilde{R}_\rho| < \infty$ when $\alpha \geq 1$. These details are summarized later in Lemma 6.

Since $D_\rho(h; \theta)$ is featured in the objective functions used to define both $R_\rho$ and $\tilde{R}_\rho$, one is naturally interested in the derivative of the map $\theta \mapsto D_\rho(h; \theta)$. The following lemma shows that we can differentiate under the integral without needing any additional conditions beyond those required for finiteness.
Lemma 5. Let $\rho_\sigma$ be as in Lemma 2. Assume that the random loss $L(h)$ is $\mathcal{F}$-measurable in general, and that $E_\mu|L(h)| < \infty$ holds whenever $1 < \alpha \leq 2$. It follows that the first two derivatives are $\mu$-integrable, namely that

$$|E_\mu \rho'_\sigma(L(h) - \theta)| < \infty, \quad |E_\mu \rho''_\sigma(L(h) - \theta)| < \infty$$

(19)

for any $\theta \in \mathbb{R}$. Furthermore, the function $\theta \mapsto D_\rho(h; \theta)$ is twice-differentiable on $\mathbb{R}$, and satisfies the Leibniz integration property for both derivatives, that is

$$\frac{d}{d\theta} D_\rho(h; \theta) \bigg|_{\theta = u} = -E_\mu \rho'_\sigma(L(h) - u), \quad \frac{d^2}{d\theta^2} D_\rho(h; \theta) \bigg|_{\theta = u} = E_\mu \rho''_\sigma(L(h) - u)$$

(20)

for any $u \in \mathbb{R}$.\(^{13}\)

With first-order information about the expected dispersion in hand, one can readily obtain conditions under which the risks $R_\rho$ and $\tilde{R}_\rho$ are determined by a finite “optimal threshold.”

Lemma 6. Following the setup of Lemma 3, let $D_\rho(h; \theta) < \infty$ for all $\theta \in \mathbb{R}$. Then, for any choice of $-\infty \leq \alpha \leq 2$ and $\sigma > 0$, there exists a finite solution $\theta^*_h \in \mathbb{R}$ such that $D_\rho(h) = D_\rho(h; \theta^*_h)$. For any choice of $\eta \geq 0$, we thus have

$$R_\rho(h) = E_\mu L(h) + \eta D_\rho(h; \theta^*_h).$$

(21)

As for the threshold risk $\tilde{R}_\rho$, if $\alpha > 1$, then for any choice of $\sigma > 0$ and $\tilde{\eta} \in \mathbb{R}$, there exists a finite optimal threshold $\tilde{\theta}^*_h \in \mathbb{R}$ such that

$$\tilde{R}_\rho(h) = D_\rho(h; \tilde{\theta}^*_h) + \tilde{\eta} \tilde{\theta}^*_h.$$  

(22)

In the case of $\alpha = 1$, we have that (22) holds if and only if $|\tilde{\eta}| < 1/\sigma$. More generally, if $\alpha < 1$, then it follows that $\tilde{R}_\rho(h) = -\infty$. Finally, in the special case of $1 \leq \alpha \leq 2$, the optimal thresholds in (21) and (22) are unique.

Thus far, we have considered properties assuming a random loss $L(h)$ given a fixed candidate $h$, but more generally, we will have a set of random losses $\mathcal{L}$ that typically is constrained by our model (e.g., $\mathcal{L} = \{L(h) : h \in \mathcal{H}\}$ under hypothesis class $\mathcal{H}$). One can then naturally consider both risks and expected dispersions as functions defined on a set of random variables; let us modify our notation to make this explicit. First, the expected deviations are re-written as

$$D_\rho(L; \theta) := E_\mu \rho_\sigma(L - \theta), \quad L \in \mathcal{L}.$$  

(23)

Next, the solutions in Lemma 6 are generalized in the same way, as

$$\theta^*_{L} \in \text{arg min} \{D_\rho(L; \theta) : \theta \in \mathbb{R}\}, \quad \tilde{\theta}^*_L \in \text{arg min} \{D_\rho(L; \theta) + \tilde{\eta}\theta : \theta \in \mathbb{R}\}.$$  

(24)

Finally, the risks are re-written as

$$R_\rho(L) := E_\mu L + \eta D_\rho(L; \theta^*_L), \quad \tilde{R}_\rho(L) := \tilde{\eta} \tilde{\theta}^*_L + D_\rho(L; \tilde{\theta}^*_L).$$  

(25)

With this notation in hand, let us consider some basic properties of the underlying dispersions, thresholds, and risks.

\(^{13}\)Let us emphasize that $\rho'_\sigma$ and $\rho''_\sigma$ denote the first and second derivatives of $x \mapsto \rho(x/\sigma; \alpha)$, which differ from $\rho'(x/\sigma; \alpha)$ and $\rho''(x/\sigma; \alpha)$ by a $\sigma$-dependent factor; see §B.3 for details.
Lemma 7. Let $\mathcal{L}$ be such that the assumptions of Lemma 6 hold for each $L \in \mathcal{L}$. Then, both dispersion quantities are translation-invariant, i.e., for any $c \in \mathbb{R}$ we have

$$D_\rho(L + c; \theta^*_L) = D_\rho(L; \theta^*_L), \quad D_\rho(L + c; \tilde{\theta}^*_L) = D_\rho(L; \tilde{\theta}^*_L).$$

They are also both non-negative; more precisely, $D_\rho(L; \theta^*_L) \geq 0$ in general, and $D_\rho(L; \tilde{\theta}^*_L) = 0$ if and only if $L$ is constant. On the other hand, $D_\rho(L; \tilde{\theta}^*_L) > 0$ for any $L$, even if constant.

These thresholds are translation-equivariant, in that we have

$$\theta^*_{L + c} = \theta^*_L + c, \quad \tilde{\theta}^*_{L + c} = \tilde{\theta}^*_L + c$$

for any choice of $c \in \mathbb{R}$. In the special case of $1 \leq \alpha \leq 2$, both of these thresholds are monotonic in the sense that whenever $L_1, L_2 \in \mathcal{L}$ and $L_1 \leq L_2$ almost surely, we have

$$\theta^*_{L_1} \leq \theta^*_{L_2} \quad \text{and} \quad \tilde{\theta}^*_{L_1} \leq \tilde{\theta}^*_{L_2}.$$ 

Furthermore, the maps $L \mapsto R_\rho(L), \ L \mapsto \tilde{R}_\rho(L)$, and $L \mapsto D_\rho(L; \theta^*_L)$ are all convex on $\mathcal{L}$.

Let us briefly discuss the properties described in Lemma 7 with a bit more context. One of the best-known classes of risk functions is that of coherent risks [2], typically characterized by properties of convexity, monotonicity, translation equivariance, and positive homogeneity [48]. Our general notion of “dispersion” is often referred to as “deviation” in the risk literature, and the properties of translation invariance, sub-linearity (implying convexity), non-negativity, and definiteness (i.e., zero only for constants) allow one to establish links between deviations and coherent risks [47]. In general, the risks $R_\rho$ and $\tilde{R}_\rho$ that we study take us outside of this traditional framework, while still maintaining lucid connections. More precisely, note that while our risks are not themselves monotonic nor translation-equivariant, these properties are instead satisfied by the optimal thresholds $L \mapsto \theta^*_L$ and $L \mapsto \tilde{\theta}^*_L$, underlying each risk. While the dispersion terms $D_\rho(L; \theta^*_L)$ and $D_\rho(L; \tilde{\theta}^*_L)$ in each risk are not “general deviations” in that they are not in general sub-linear, they do satisfy the critical properties of translation invariance, non-negativity, and definiteness. In addition, when the underlying $\rho$ is convex, this convexity is inherited by both risks and the M-dispersion.

4 Learning theory

In the previous section, our focus was on risk functions used for testing, which are ideal quantities in that they are defined using an unknown data distribution. The next point of interest is linking the risk functions $R_\rho$ and $\tilde{R}_\rho$ (with $\rho$ from the Barron class) that we have studied thus far to practical feedback generation and learning algorithms, that is, the design of training procedures. The purpose of this section is to capture the risk families of interest using standard analytical machinery for stochastic gradient-based algorithms.

To maintain a narrative that is easy to follow, we break this section into several sub-sections. First, in §4.1 we introduce generalized objectives given in terms of both the parameter $h$ and threshold $\theta$, and discuss how learning performance guarantees are inevitably local, and will be measured in terms of the stationarity of an iterative learning algorithm. In §4.2 we spell out the details on obtaining unbiased stochastic gradients, and establish smoothness properties for the generalized objectives in §4.3. We describe a concrete gradient-based learning algorithm in §4.4, providing formal stationarity guarantees for this procedure in §4.5, and a discussion of technical limitations and directions forward in §4.6.
4.1 Measuring performance

Recalling the definitions (3) and (4), both $R_\rho$ and $\tilde{R}_\rho$ are defined using random dispersion measured about an optimal threshold. Let $\mathcal{H}$ denote the hypothesis class from which the learning algorithm must select a candidate. Instead of directly minimizing these risks over $\mathcal{H}$, arguably the most natural approach is to slightly broaden the learning problem, treating the threshold $\theta \in \mathbb{R}$ as a parameter to be determined alongside $h \in \mathcal{H}$ during the learning process. More explicitly, we introduce the following intermediate objectives:

$$r_\rho(h, \theta) := E_\mu L(h) + \eta D_\rho(h; \theta)$$
$$\tilde{r}_\rho(h, \theta) := D_\rho(h; \theta) + \tilde{\eta}\theta.$$  \hspace{1cm} (29)

Comparing (3) and (4) with (29) and (30), clearly we have $R_\rho(h) \leq r(h, \theta)$ and $\tilde{R}_\rho(h) \leq \tilde{r}_\rho(h, \theta)$ for any choice of $(h, \theta) \in \mathcal{H} \times \mathbb{R}$. If $R_\rho(\cdot)$ and $\tilde{R}_\rho(\cdot)$ respectively take their minimum on $\mathcal{H}$ at $h^*$ and $\tilde{h}^*$, and the losses $L(h^*)$ and $L(\tilde{h}^*)$ satisfy the conditions given in Lemma 3, then via Lemma 6 there exist pairs $(h^*, \theta^*)$ and $(\tilde{h}^*, \tilde{\theta}^*)$ such that

$$r_\rho(h^*, \theta^*) = R_\rho^*$$
$$\tilde{r}_\rho(\tilde{h}^*, \tilde{\theta}^*) = \tilde{R}_\rho^*$$  \hspace{1cm} (31)

where $R_\rho^* := \inf\{R_\rho(h) : h \in \mathcal{H}\}$ and $\tilde{R}_\rho^* := \inf\{\tilde{R}_\rho(h) : h \in \mathcal{H}\}$. There is no gap between the optimal function values; solving the joint objectives (29) and (30) is clearly sufficient for solving the original objectives.

One would ideally like to use the excess risks $R_\rho - R_\rho^*$ and $\tilde{R}_\rho - \tilde{R}_\rho^*$ to express algorithm performance guarantees, but unfortunately this is not in general possible due to a lack of convexity. While we know that the function $\rho(\cdot) = \rho(\cdot; \alpha)$ is convex on $\mathbb{R}$ for $1 \leq \alpha \leq 2$, since $\rho$ is not monotonic on $\mathbb{R}$ for any choice of $\alpha$, regardless of the convexity of $h \mapsto L(h)$, in general the map $h \mapsto D_\rho(h; \theta)$ cannot be convex.\footnote{As a simple illustrative example, consider the function $f(x) := (x^2 - \theta)^2$, where we have $f''(x) = 12x^2 - 4\theta$. For any $\theta > 0$, clearly this function $f$ is concave near zero, and convex elsewhere.} In general, the convexity of both $r_\rho(\cdot)$ and $\tilde{r}_\rho(\cdot)$ cannot be assumed; this holds for all choices of $\eta > 0$, $\tilde{\eta} \in \mathbb{R}$, $\sigma > 0$, and $-\infty \leq \alpha \leq 2$. Since convergence to a global minimum cannot be guaranteed in general, it is typical to state guarantees in terms of stationarity of the learning algorithm, i.e., finding a point from which an iterative learning algorithm does not change much, even if it were given ideal information about the objective function. A concrete example of a general-purpose stationarity indicator for gradient-based algorithms is (46) given in §4.5.

4.2 Unbiased stochastic feedback

Since $r_\rho(h, \theta)$ and $\tilde{r}_\rho(h, \theta)$ cannot be directly observed, all we have is noisy (stochastic) estimates of their true values. This amounts to a transformation of the base loss $L(h)$ in accordance with the underlying risk of interest; we denote these modified losses as

$$L_\mu(h, \theta) := L(h) + \eta \rho_{\sigma}(L(h) - \theta)$$
$$\tilde{L}_\mu(h, \theta) := \rho_{\sigma}(L(h) - \theta) + \tilde{\eta}\theta.$$  \hspace{1cm} (32)

$$\tilde{L}_\mu(h, \theta) := \rho_{\sigma}(L(h) - \theta) + \tilde{\eta}\theta.$$  \hspace{1cm} (33)

Since our assumption from the start is that the random base loss $L(h)$ can be sampled, and considering the fact that $\eta$, $\tilde{\eta}$, $\rho$, $\sigma$, and $\theta$ are all known quantities, it follows that we can also readily sample the modified losses $L_\mu$ and $\tilde{L}_\mu$. This gives us some rudimentary feedback which is (trivially) guaranteed to be unbiased, since $E_\mu L_\mu = r_\rho$ and $E_\mu \tilde{L}_\mu = \tilde{r}_\rho$ follows from definitions.
For many learning algorithms, however, this alone is insufficient, since unbiased first-order feedback is often required. We already know conditions for the map \( \theta \mapsto D_\rho(h; \theta) \) to be continuously differentiable on \( \mathbb{R} \) and to satisfy the Leibniz integration property, via Lemma 5. This means that writing the partial derivatives with respect to \( \theta \) as

\[
\partial_\theta L_\rho(h, \theta) := -\eta \rho'_\sigma(L(h) - \theta) \\
\partial_\theta \tilde{L}_\rho(h, \theta) := \tilde{\eta} - \rho'_\sigma(L(h) - \theta)
\]

(34) (35)

it follows from Lemma 5 that \( E_\mu \partial_\theta L_\rho = \partial_\theta r_\rho \) and \( E_\mu \partial_\theta \tilde{L}_\rho = \partial_\theta \tilde{r}_\rho \). It thus remains to consider taking partial derivatives with respect to \( h \in \mathcal{H} \). This inevitably requires us to look at the smoothness of \( h \mapsto L(h) \), and thus depends on how we measure distance on \( \mathcal{H} \). Fortunately, the desired property holds under quite general settings, for any choice of \(-\infty \leq \alpha \leq 2\), as we confirm in the following lemma.

**Lemma 8.** Let \( \mathcal{U} \) be an open subset of any metric space such that \( \mathcal{H} \subset \mathcal{U} \). Let the base loss map \( h \mapsto L(h) \) be Fréchet differentiable on \( \mathcal{U} \) (\( \mu \)-almost surely), with gradient denoted by \( L'(h) \) for each \( h \in \mathcal{U} \). Let partial derivatives of the modified losses be denoted by

\[
\partial_h L_\rho(h, \theta) := (1 + \eta \rho'_\sigma(L(h) - \theta)) L'(h) \\
\partial_h \tilde{L}_\rho(h, \theta) := \rho'_\sigma(L(h) - \theta)L'(h)
\]

(36) (37)

and assume that the loss distribution and \( \mathcal{H} \) are such that

\[
E_\mu \left[ \sup_{h \in \mathcal{H}} \|L(h)L'(h)\| \right] < \infty.
\]

(38)

It then follows that \( E_\mu \partial_h L_\rho = \partial_h r_\rho \) and \( E_\mu \partial_h \tilde{L}_\rho = \partial_h \tilde{r}_\rho \) hold on \( \mathcal{H} \).

Taking the above points together, under the assumptions of Lemmas 5 and 8, we can obtain unbiased stochastic gradients in the sense that the equalities

\[
E_\mu(\partial_h L_\rho, \partial_\theta L_\rho) = r_\rho' \quad \text{and} \quad E_\mu(\partial_h \tilde{L}_\rho, \partial_\theta \tilde{L}_\rho) = \tilde{r}_\rho'
\]

(39)

are valid on \( \mathcal{H} \times \mathbb{R} \).

### 4.3 Smoothness of the objective

When the objectives \( r_\rho \) and \( \tilde{r}_\rho \) are sufficiently smooth, there are well-known analytical techniques which provide conditions for stochastic gradient-based methods to converge to a stationary point. Formally, we will require that the maps \( (h, \theta) \mapsto r_\rho'(h, \theta) \) and \( (h, \theta) \mapsto \tilde{r}_\rho'(h, \theta) \) are Lipschitz continuous. In checking whether such smoothness indeed holds, assuming we have unbiased first-order stochastic feedback as in (39), we will always have to deal with terms of the form \( E_\mu \left[ \rho'_\sigma(L(h) - \theta)L'(h) \right] \) (recall (36) and (37)). Defining \( f(h, \theta) := \rho'_\sigma(L(h) - \theta)L'(h) \) for readability, and considering the function difference at two arbitrary points \( (h_1, \theta_1) \) and \( (h_2, \theta_2) \), first note that

\[
f(h_1, \theta_1) - f(h_2, \theta_2) = \rho'_\sigma(L(h_1) - \theta_1) [L'(h_1) - L'(h_2)] + [\rho'_\sigma(L(h_1) - \theta_1) - \rho'_\sigma(L(h_2) - \theta_2)] L'(h_2).
\]

(40)

In the case of \( \rho(\cdot) = \rho(\cdot; \alpha) \) from the Barron class (5), when \(-\infty \leq \alpha \leq 1\), we have that \( \rho' \) is both bounded (\( \|\rho'\|_\infty < \infty \)) and Lipschitz continuous on \( \mathbb{R} \) (see Lemma 2). This means that
all we need in order to control \( E_\mu A + E_\mu B \) for \( L \) to be smooth (for control of \( A \)) and for \( L'(\cdot) \) to have a norm bounded over \( H \) (for control of \( B \)); see §A.3.1 for more details. Things are more difficult in the case of \( 1 < \alpha < 2 \), since the dispersion function \( \rho \) is not (globally) Lipschitz, meaning that \( \|\rho\|_\infty = \infty \). Even if \( L \) is smooth, when the threshold parameter is left unconstrained, it will always be possible for \( \|E_\mu A\| \to \infty \) as \( |\theta_1| \to \infty \), impeding smoothness guarantees for \( r_\rho(\cdot) \) and \( \tilde{r}_\rho(\cdot) \) in this setting.

Let us proceed by distilling the preceding discussion into a set of concrete conditions that are sufficient to make the objectives \( r_\rho \) and \( \tilde{r}_\rho \) amenable to standard analysis techniques for stochastic gradient-based algorithms. For readability, we write \( \|L'\|_H := \sup \{\|L'(h)\| : h \in H\} \).

**A1. Moment bound for loss gradient.** For any choice of \( h_1, h_2 \in H \), \( 0 < c < 1 \), and \( k \in \{1, 2\} \), the loss \( L \) is differentiable at \( ch_1 + (1 - c)h_2 \), and satisfies

\[
E_\mu \left( \sup_{0 < c < 1} \|L'(ch_1 + (1 - c)h_2)\| \right)^k \leq E_\mu \|L'\|_H^k < \infty.
\]

**A2. Loss is smooth in expectation.** There exists \( 0 < \lambda_1 < \infty \) such that for any choice of \( h_1, h_2 \in H \), we have \( E_\mu \|L'(h_1) - L'(h_2)\| \leq \lambda_1 \|h_1 - h_2\| \).

**A3. Dispersion is Lipschitz and smooth.** The function \( \rho \) is such that \( \|\rho\|_\infty < \infty \), and there exists \( 0 < \lambda_2 < \infty \) such that \( |\rho'(x_1) - \rho'(x_2)| \leq \lambda_2 |x_1 - x_2| \) for all \( x_1, x_2 \in \mathbb{R} \).

If \( H \) is a convex set, then the first inequality in A1 holds trivially. Note that under A2, the right-hand side of (41) will be finite for \( k = 1 \) whenever \( H \) has bounded diameter and \( E_\mu \|L'(h)\| < \infty \) for some \( h \in H \). As for A3, all the requirements are clearly met by the Barron class with \(-\infty \leq \alpha \leq 1 \). We have introduced these conditions for a reason; they are sufficient for obtaining smoothness guarantees for the intermediate objectives (29) and (30), as summarized in the lemma below.

**Lemma 9.** Let the conditions A1, A2, and A3 hold. Then, the intermediate risks \( r_\rho \) and \( \tilde{r}_\rho \) are smooth on \( H \times \mathbb{R} \) in the sense that

\[
\|r_\rho'(h_1, \theta_1) - r_\rho'(h_2, \theta_2)\| \leq \left( \lambda_1 + \left( \frac{\eta \lambda_2}{\sigma} \right) E_\mu \|L'\|_H \right) (\|h_1 - h_2\| + |\theta_1 - \theta_2|),
\]

\[
\|\tilde{r}_\rho'(h_1, \theta_1) - \tilde{r}_\rho'(h_2, \theta_2)\| \leq \left( \frac{\lambda_5}{\sigma} + \frac{\lambda_2}{\sigma^2} E_\mu \|L'\|_H \right) (\|h_1 - h_2\| + |\theta_1 - \theta_2|)
\]

for any choice of \( h_1, h_2 \in H \) and \( \theta_1, \theta_2 \in \mathbb{R} \). Here the factor \( \lambda_5 \) is defined \( \lambda_5 := \lambda_3 + \lambda_4 \), where

\[
\lambda_3 := \left( \frac{\lambda_2}{\sigma} \right) E_\mu \|L'\|_H^2 + \sup_{h \in H} E_\mu \|L'(h)\|, \quad \lambda_4 := \lambda_1 \|\rho\|_\infty.
\]

Proving Lemma 9 is straightforward but somewhat tedious. Detailed computations as well as a direct proof are organized in §A.3.1 for easy reference.

**Remark 10 (Norm on product spaces).** In Lemma 9 we have to deal with norms on product spaces, and in all cases we just use the traditional choice of summing the norms of the constituent elements, i.e., on \( H \times \mathbb{R} \), we have \( \|(h, \theta)\| := \|h\| + |\theta| \). Similarly, as is evident from (39), \( r_\rho'(h, \theta) \) and \( \tilde{r}_\rho'(h, \theta) \) are both represented as pairs of linear functionals; their norm is defined as the sum of the norms of these two constituent functionals.
4.4 A concrete learning algorithm

With conditions in hand to ensure smoothness, it is fairly straightforward to “plug in” this smoothness to standard analytical procedures for stochastic gradient-based algorithms to obtain lucid performance guarantees. For arbitrary sequences \((h_t)\) and \((\theta_t)\), the basic stochastic gradient feedback sequence we use is denoted as \((G_t)\), and defined as follows:

\[
G_t := \begin{cases} 
L'_\rho(h_t, \theta_t), & \text{if minimizing } r_ho \\
\tilde{L}'\rho(h_t, \theta_t), & \text{if minimizing } \tilde{r}_\rho.
\end{cases}
\]  

(42)

The modified loss gradients are respectively \(L'_\rho = (\partial_h L_\rho, \partial_\theta L_\rho)\) and \(\tilde{L}'\rho = (\partial_h \tilde{L}_\rho, \partial_\theta \tilde{L}_\rho)\), with partial derivatives precisely as defined in §4.2. We denote sub-sequences as \(G[t] := (G_1, \ldots, G_t)\) for all \(t > 0\). We use a single symbol \(G_t\) for random gradients under both \(r_\rho\) and \(\tilde{r}_\rho\); this is in order to eliminate redundancy, emphasizing key facts with concise notation. Before giving a concrete learning algorithm, we organize a few additional technical conditions.

B1. \(H\) is a subset of some Hilbert space, and \(H \times \mathbb{R}\) is a closed convex set.

B2. The feedback (42) satisfies \(E[G_t \mid G[t-1]] = E \mu G_t\) and (39) for all \(t > 1\).\(^{15}\)

In general, the gradient \(G_t\) is a linear functional taking arguments from \(H \times \mathbb{R}\), but since all Hilbert spaces are reflexive Banach spaces, this can be uniquely identified with an element of \(H \times \mathbb{R}\), for which we use the same notation (namely, \(G_t\)). The learning algorithm we consider is given by updates of the form

\[
(h_{t+1}, \theta_{t+1}) = \text{Step}_{H \times \mathbb{R}}[(h_t, \theta_t); \mathcal{G}_t, \gamma_t]
\]  

(43)

with \(\mathcal{G}_t\) denoting the \(m\)-sized “mini-batch” average

\[
\mathcal{G}_t := \frac{1}{m} \sum_{i=1}^{m} G_{t,i}
\]  

(44)

where \(G_{t,1}, \ldots, G_{t,m}\) are all independent copies of \(G_t\). The key sub-routine underlying the update (43) is defined by

\[
\text{Step}_U[u; g, \gamma] := \Pi_U[u - \gamma g]
\]  

(45)

where \(U\) is an arbitrary set, \(u, g\) are arbitrary vectors, \(\gamma\) is an arbitrary scalar, and \(\Pi_U[\cdot]\) denotes projection to \(U\). As such, the controllable parameters of the learning algorithm (43) are the non-negative step-sizes \((\gamma_t)\) and the mini-batch size \(m > 0\).

4.5 Stationarity guarantees

Next we describe performance guarantees that hold for the learning algorithm (43) described in the previous sub-section. We assume that the recursion given by (43) is started by some initial point \((h_0, \theta_0) \in H \times \mathbb{R}\). To measure performance, recall from our discussion in §4.1 that we will focus on showing that given enough data, on average, the learning algorithm of interest finds a point which is nearly stationary. More precisely, the degree of stationarity at any time step \(t > 0\) will be measured using

\[
\Delta_t := \frac{1}{\gamma_t} \langle (h_t, \theta_t) - \text{Step}_{H \times \mathbb{R}}[(h_t, \theta_t); \mathcal{E}_\mu G_t, \gamma_t] \rangle.
\]  

(46)

\(^{15}\)The expectation on the left-hand side is with respect to the joint distribution of \(G[t]\) conditioned on \(G[t-1]\).
Note that $\Delta_t$ is an ideal quantity depending on the unknown expected gradients $E_\mu G_t$, representing the re-scaled difference that would arise if it were possible to update using the true gradient (of $r_\rho$ or $\tilde{r}_\rho$). Intuitively, this distance will become small if we are close to a stationary point. From a technical standpoint, while it can be difficult to control the norm of $r_\rho$ and $\tilde{r}_\rho$ directly, it is much easier to control the norm of $\Delta_t$. When $\Delta_t$ is sufficiently small, the approximation $\Delta_t \approx r_\rho^*(h_t, \theta_t)$ (or $\Delta_t \approx \tilde{r}_\rho^*(h_t, \theta_t)$) becomes increasingly sharp. We remark that the use of $\Delta_t$ as a performance indicator is standard in the literature for non-convex constrained stochastic optimization [21]. Given the computational procedure and technical conditions laid out in §4.2–§4.4, the notion of “nearly stationary on average” is made precise in the following theorem.

**Theorem 11.** Run the process (43) for $n$ iterations, with mini-batches of size $m$. Let $T$ be a random variable with support $\{0, \ldots, n - 1\}$, and probabilities defined as

$$P\{T = t\} = \frac{\gamma_t - \lambda \gamma_t^2}{\sum_{i=0}^{n-1}(\gamma_i - \lambda \gamma_i^2)}.$$ 

Let the step-size parameters be fixed as $\gamma_t = 1/(2\lambda)$, where $\lambda$ is the smoothness coefficient from Lemma 9 (depending our choice of $r_\rho$ versus $\tilde{r}_\rho$). Then under A1–A3 and B1–B2, we have

$$E||\Delta_T||^2 \leq \frac{c_0 \lambda}{n} + \left(\sup_{h \in \mathcal{H}}E_\mu ||L'(h) - E_\mu L'(h)||^2\right) \frac{6}{m}$$

where expectation on the left-hand side is taken over all the stochastic gradients and the choice of $T$, and $c_0$ is the initialization error (either $r_\rho(h_0, \theta_0) - r_\rho^*$ or $\tilde{r}_\rho(h_0, \theta_0) - \tilde{r}_\rho^*$).

The message of Theorem 11 is that we have a concrete procedure for which we are guaranteed in expectation to be able to find a point which is arbitrarily “good” in terms of stationarity. That is, either the true gradient norm (either $||r_\rho||$ or $||\tilde{r}_\rho||$) is small, or we are near the boundary of $\mathcal{H} \times \mathbb{R}$ and a gradient-based update is essentially orthogonal to $\mathcal{H} \times \mathbb{R}$, meaning that the projection takes us back where we started. Regarding the sample complexity, say we are constrained to $N$ data points in total. If we set $n = m = \sqrt{N}$, then in total we use $n \times m = N$ points and achieve an upper bound of $O(1/\sqrt{N})$, meaning that $O(\varepsilon^{-4})$ samples are sufficient to ensure that $\sqrt{E||\Delta_T||^2} \leq \varepsilon$.

### 4.6 Moving beyond the standard approach

The conditions and computational procedure adopted in the preceding sub-sections can be considered an application of well-established insights from stochastic optimization to a new set of learning tasks borne from the risks (3) and (4). That said, there still remains significant room for improvement. The main limitations to the analysis above are as follows: the assumption of a Hilbert space (from B1), random stopping (random variable $T$ in Theorem 11), guarantees given only in expectation rather than with high probability, and constraints on the Barron class requiring $\alpha \leq 1$. We briefly expand on these points.

The assumption of a Hilbert space is common, and makes the analysis much easier when the squared norm can be written as an inner product and decomposed in a convenient fashion. While optimization on arbitrary Banach spaces is challenging even in the convex case, recent work from Cutkosky and Mehta [14] highlights an effective proof technique for stochastic gradient-based algorithms when working with some important classes of Banach spaces. Furthermore, these insights apply to smooth, non-convex objectives, yielding guarantees that hold with high probability, without having to resort to random stopping. The computational
procedure is more involved and assumes an unconstrained parameter space, but the key strategy of combining gradient clipping and normalization is a promising approach for designing a new procedure that goes well beyond the guarantees of Theorem 11. As for the question of constraints on shape parameter $\alpha$, obviously this problem can be addressed in a brute-force way by constraining the choice of $(h, \theta)$ to a bounded subset of $\mathcal{H} \times \mathbb{R}$, forcing $\sup_{h \in \mathcal{H}} |L(h)|$ to have a bounded support (implying that $\rho'_\alpha (L(h) - \theta)$ is bounded almost surely), and finally modifying $\lambda_4$ in Lemma 9 accordingly, but such conditions are very restrictive. When both $\rho'$ and $L(h)$ are unbounded, it will be important to have some idea of the magnitude of the minimizers of $\theta \mapsto r(h, \theta)$, and how this can change as $h$ ranges over $\mathcal{H}$. Prior knowledge of such magnitudes allows one to focus on conditions for the loss distribution when trying to control the term $A$ from (40). A cleaner, more satisfying solution would involve characterizing a high-probability event in which the sequence $(\theta_t)$ cannot deviate arbitrarily far from a stationary point.

**Remark 12 (Weak convexity).** Even though the objective $r(\cdot)$ is differentiable and the gradient $r'(\cdot)$ is locally Lipschitz, for the setting of $1 < \alpha \leq 2$, we cannot assume that $(h, \theta) \mapsto r(h, \theta)$ is smooth on the whole of $\mathcal{H} \times \mathbb{R}$, and thus one might be inclined to treat the problem as non-convex and non-smooth from the start. Among such problems, when we can show a form of weak convexity, it is often possible to provide guarantees for stochastic gradient-based algorithms which are analogous to the smooth case, but without requiring a smooth objective. Unfortunately, such arguments require $\rho$ to be both convex and Lipschitz on $\mathbb{R}$. Convexity fails when $\alpha < 1$, and Lipschitz continuity fails when $\alpha > 1$. As such, only the special case of $\alpha = 1$ allows for weak convexity arguments.

## 5 Empirical analysis

Taking a high-level perspective, the most basic point of inquiry underlying this research is that of the relationship between (base) loss distribution properties and the choice of risk function. This is the “static” side of the problem; given a loss distribution with certain properties, which risk classes are sensitive to those properties? There is also a “dynamic” side to the problem, since in any iterative learning algorithm, the loss distribution is constantly changing with each parameter update step. Given a hypothesis class, data-generating process, and learning algorithm, how does the choice of risk function (and resulting loss transformation) impact the process and outcomes of learning? These are big questions, but to provide some concrete initial answers, in this section we describe results from a series of numerical experiments designed to shed light on the similarities and differences of the main risk classes of interest from §2.

In order to ensure this section is easy to follow, we break the exposition into multiple sub-sections, and since we include several large figures, we have reserved the last sub-section §5.5 for all the figures relevant to this section. We start by looking at risk characteristics under a wide variety of simulated loss distributions in §5.1, and complement this in §5.2 with some examples of real loss distributions that arise when initializing typical models under traditional datasets. We then shift our focus to learning algorithms, and begin in §5.3 by paying particular attention to the role that risk design plays when the data includes a small fraction of extreme outliers. Finally, we look at the loss trajectories and final base loss distributions that arise when tackling well-known classification datasets in §5.4.

\footnote{See for example previous work by Holland [26], which was restricted to a special class of risks built using convex Lipschitz functions that are amenable to new proof techniques utilizing weak convexity [15].}
Software  In order to ensure that the empirical analysis is transparent and reproducible, we have created a public repository (https://github.com/feedbackward/bdd) of software and Jupyter notebooks that can be used to re-create all of the experimental results and figures which appear in this paper.

5.1 Simulation-based risk class analysis

To begin, we simulate a variety of base loss distributions by taking large samples from various well-known parametric distributions, and take a close look at how each risk class behaves. Intuitively, each distinct distribution studied in this sub-section can be interpreted as the distribution of $L(h)$ for different choices of $h \in \mathcal{H}$. We will see the kinds of distributions that actually arise under typical models and datasets in subsequent subsections.

5.1.1 Experimental setup

Letting $L$ denote the random loss we are simulating, we specify a parametric distribution for $L$, from which we take an independent sample $\{L_1, \ldots, L_m\}$.\(^{17}\) In all cases, we center the true distribution such as $E_{\mu} L = 0$.\(^{18}\) Based on this sample, we estimate the following risk classes: M-risk from (3) and T-risk from (4) under the Barron class (5), CVaR risk in (14), entropic risk from (13), and $\chi^2$-DRO risk (16) as defined in §2.4. Estimation is done by replacing the true expectation $E_{\mu}$ in each of the risk definitions by the empirical mean, and numerically solving for the optimal threshold $\theta$ that characterizes each risk.\(^{19}\) We record both risk values as well as optimal threshold values. To ensure that we capture properties of the true underlying distribution, we take a large sample size of $m = 10^4$.

We stated earlier that we want to evaluate the “behavior” of each risk class. By this we mean that given a pre-fixed loss distribution, we want to see how different choices within each risk function class changes the risk value that is returned. For M-risk here, this will amount to modifying $\alpha$ and $\sigma$. For T-risk, we will modify $\hat{\eta}$ in addition to $\alpha$ and $\sigma$. More precisely, when modifying $\alpha$ over $[1, 2]$, we set $\sigma = 0.5$, $\eta = 1.0$, $\tilde{\eta} = 0.99$.\(^{20}\) When modifying $\sigma$ over $[0.05, 1.5]$, we fix $\alpha = 1.0$, $\eta = 1.0$ and $\tilde{\eta} = 1/(1.5) - 0.01$. When modifying $\tilde{\eta}$ over $[-1, 1]$, we fix $\alpha = 1.0$ and $\sigma = 0.99$. For CVaR we modify the underlying quantile level $\beta$ over $[0.025, 0.975]$. For entropic risk we modify the parameter $\gamma$ over $[0.01, 2.0]$. Finally, for $\chi^2$-DRO, we re-parameterize the upper bound from (17) as $a = ((1 - \tilde{a})^{-1} - 1)^{2}/2$, and modify $\tilde{a}$ ranging over $[0.025, 0.975]$. Thus, for each distribution being studied, with an $m$-sized sample in hand, we compute risks using many risk functions from each class (using the exact same sample). This lets us study both within-class and between-class trends.

5.1.2 Results and discussion

We start with Figures 3–4, in which we consider two families (Exponential and Gamma) that are both asymmetric and unbounded, but differ in their dispersion and tail properties. In addition to comparing across these two families, we also look at how flipping the underlying distribution (replacing $L$ with $-L$) impacts the resulting risk values. In these figures, we plot both risk values (dashed black curves) and optimal thresholds (solid black curves), and note

\(^{17}\)The distribution settings were fixed in an arbitrary manner, before any testing took place.
\(^{18}\)We remark that this is just for ease of numerical comparison across distributions. The overall insights do not change under shifts.
\(^{19}\)We use ``minimize_scalar`` from SciPy, with bounded solver type, and valid brackets set manually.
\(^{20}\)The choice to fix $\sigma = 0.5$ was based on a comparison of three choices: $\sigma = 0.1, 0.5, 1.0$, of which $\sigma = 0.5$ was visually confirmed to be the most suitable scale.
that under the setup described, one can interpret the difference between these two values as the “dispersion part” of each risk. For emphasis, we have colored this gap in light gray, with the exception of the entropic risk, for which both values are equal. Among the most salient features is the impact that the direction of the distribution tails has on risk values. Comparing the top and bottom rows in each figure, we see that CVaR, entropic, and \( \chi^2 \)-DRO risks all change quite dramatically, whereas the only the T-risk is invariant, sensitive to dispersion on both the upside and the downside. Another interesting point is the relative impact that the tails have on each risk. Comparing the Exponential and Gamma cases, extreme values are much farther from the majority under the Exponential distribution, but the majority of points are also more sharply concentrated than in the Gamma case. With this fact in mind, we can observe that CVaR, entropic, and \( \chi^2 \)-DRO risks are sensitive to rare but very distant values, whereas the T-risk over \( \alpha \) is less sensitive to extreme values, and more sensitive to the spread of the distribution around the median.

Next, we proceed to Figure 5, which is similar to the previous two figures, except that we now consider more distributions, this time without flipping the distribution, and we zoom out to plot a wider range of quantiles, marked for visual reference. One particularly important point is the stark difference in how each risk class deals with extremely heavy-tailed data. The T-risk exceeds the 99th percentile when the majority has a large spread (e.g., Gamma), but remains well below the extreme quantiles when most of the spread is due to a few extreme outliers (e.g., Pareto). This is contrast to the other risk classes, which have a stronger tendency to be pulled by rare outliers. We also remark that in principle, the risks being considered here can be defined for discrete distributions, and even for very simple cases such as an asymmetric Bernoulli, each risk class behaves in a rather distinct way.

Finally, in Figure 6 we take a deeper look at M-risk and T-risk under the Barron class. We look at changes in T-risk under modifications to \( \sigma \) and \( \tilde{\eta} \), and for comparison also give results for M-risk under modified \( \alpha \) and \( \sigma \). Starting with impact of \( \alpha \) on M-risk, note that the symmetry of the underlying data plays an important role on the optimal threshold; it is invariant under changes to \( \alpha \) when the data is symmetric, but varies when the data is asymmetric. This is an immediate byproduct of the symmetry of \( \rho \). Compare this with T-risk, where the optimal threshold varies with \( \alpha \) even when the data is symmetric (cf. Figure 5). Next, we can observe that the impact of scale differs in a rather stark way between M-risk and T-risk. Perhaps the most obvious difference is that (by definition) \( R_\rho \) cannot go below the mean, whereas \( \tilde{R}_\rho \) can range well above and below the mean with proper scale settings. Another point worth noting is how the optimal thresholds of each risk class change: for \( R_\rho \), moving from small to large value of \( \sigma \) always takes one “towards the mean” in the limit, though this can be both increasing and decreasing, whereas the optimal threshold in \( \tilde{R}_\rho \) always decreases as \( \sigma \) increases. Note that this monotonicity is inevitable in \( \tilde{R}_\rho \) whenever \( \alpha \geq 1 \) (note that \( \rho' \) is non-decreasing on \( \mathbb{R} \)), but this changes when \( \alpha < 1 \) due to the re-descending nature of \( \rho' \) (recalling Figure 1). Finally, the impact of \( \tilde{\eta} \) is quite lucid; \( \tilde{\eta} < 0 \) pushes the threshold up, and \( \tilde{\eta} > 0 \) pushes the threshold down, just as we would expect from the first-order condition \( E_\mu \rho'_\mu(L(h) - \theta) = \sigma \tilde{\eta} \) when \( \rho' \) is monotonically increasing on \( \mathbb{R} \), i.e., when \( \alpha \geq 1 \). Note also that with appropriate choice of \( \tilde{\eta} \), we can take the optimal threshold to any real value; this cannot be done by simply adjusting the scale \( \sigma \). Indeed, one look at the graphs in the third and fourth plots shows that \( \sigma \) and \( \tilde{\eta} \) can be used to realize a wide variety of distinct risk-threshold pairs.

\(^{21}\) See §A.1.2 for additional details.
5.2 Loss distributions at initialization

Since our focus in the previous sub-section was on simulated loss distributions, it is natural to ask what kind of loss distributions actually arise when solving real machine learning problems. In this sub-section, we introduce standard machine learning models and loss functions into the picture, and evaluate how model initialization impacts the base loss distribution. This is important in obtaining some much-needed additional perspective on the findings of §5.1, and can shed light on which of these findings we can expect to be relevant to machine learning practitioners.

To begin, we start with a binary classification task using a linear model in Figure 7. We have generated data from two Gaussian distributions, assigned positive (top left) and negative (bottom right) labels to each, and are considering two different candidates for a linear classifier on the plane, namely the red-dashed and blue-dotted lines. We then compute losses on this sample for both of these candidates using three standard loss functions: logistic loss, hinge loss, and unhinged loss, where the histogram colors correspond to the classifier candidate being evaluated. This is obviously a toy example, but it highlights how easy it is to run into highly asymmetric, heavy-tailed loss distributions in both the upward and downward directions.

Another simple example is given in Figure 8, in which we consider linear regression in one dimension, using a well-known real dataset including a small subset of outliers.22 Once again we consider two simple candidates, of identical slope (namely, zero) and differing intercept (zero and one), denoted by the red-dashed and blue-dotted lines. We then compute losses on this dataset using three loss functions commonly used in regression: squared error, Huber, and absolute error, and plot the histograms in the same way as described in the previous paragraph. Once again, we see how easy it is for asymmetric loss distributions with heavy tails to occur on both the upside and the downside.

Finally, in Figure 9 we consider some larger real-world classification datasets.23 In all of these datasets, categorical features are given a one-hot representation, and numerical features are normalized to the unit interval. These features are then passed to a randomly initialized linear model, using which we compute the (multi-class) logistic loss on both training and testing subsets (using a random 80-20 split), and it is these sets of losses whose histograms we have plotted. Random initialization sets each weight independently by drawing uniformly from $[-0.05, 0.05]$. The most obvious take-away is the high degree of diversity that we see in the loss distributions. Despite compressing all features into the unit interval, and using the exact same model and initialization procedure, we see that loss distribution properties such as asymmetry, multi-modality, dispersion, and the presence of outliers depends greatly on the underlying dataset. We take this as strong evidence that many of the differences between risk classes highlighted in §5.1 can also be reasonably expected to arise and be relevant in real-world learning settings.

5.3 Outliers and risk design

The next step in our analysis is to introduce an iterative learning algorithm into the experiment design, and to consider the impact that risk function design can have on the learning process and its outcome. In this sub-section, we focus on a simple regression task with outlying data points.

22 This is the Belgian phone call dataset [52], normalized to the unit interval.
23 See §C for a description of the datasets.
5.3.1 Experimental setup

As a well-known example that is conducive to visualization, we use the normalized Belgian phone call dataset that appeared previously in §5.2. We conduct two main experiments. First, we fix the base loss function (quadratic loss), and consider learning under the various risk classes studied in §5.1, running empirical risk minimization (ERM) using batch gradient descent for each risk, and recording the trajectories of the average base and transformed losses as we go. Second, we fix one risk function from each class, and consider the impact of modifying the underlying loss function itself, rather than the risk function. In order to limit the impact of outliers, for M-risk we set $\alpha = -\infty$, and T-risk we set $\alpha = 1.0$, both the lowest settings possible. Similarly, for CVaR risk we set $\beta = 0.5$, for entropic risk we set $\gamma = 1.0$, and for $\chi^2$-DRO we set $\tilde{a} = 0.25$. As a general representation, our random base losses will take the form $L(h) = (h(X) - Y)^c/c$ for $1 \leq c \leq 2$, where $h(X) = w_1X + w_0$ since this is one-dimensional linear regression. For the first round of experiments, $c$ is fixed to $c = 2$ while the risk functions are modified, whereas in the second round of experiments, the risk functions are fixed with the parameters just mentioned, while $c$ is modified.

Regarding the learning algorithm details, for each risk class, we are running ERM jointly in the candidate $h$ and the threshold $\theta$ for each risk class. For M-risk and T-risk, this amounts to simply replacing $E_\mu$ with the empirical expectation in the definitions of $r_\rho$ from (29) and $\tilde{r}_\rho$ from (30), respectively. For CVaR and entropic risks, since these are both OCE risks taking the form (11), we are just running ERM on $(h, \theta) \mapsto \theta + E_\mu \psi(L(h) - \theta)$ with the concrete settings of $\phi$ described in §2.4. The exact same approach goes for $\chi^2$-DRO risk, this time applied to the objective in $(h, \theta)$ that appears in (18). In all cases, we initialize $h$ as given in the previous paragraph by setting $w_0 = 1$ and $w_1 = 0$, and the initial value of $\theta$ is determined by a random uniform sample from the interval $[0, 0.05]$. For each risk function considered, we run a single trial, using full-batch gradient descent with a fixed step size of $0.005$ for 15000 iterations. We record the empirical mean of the base loss and the modified loss at each step.

5.3.2 Results and discussion

Representative results for the first round of experiments is contained in Figures 10–11. Changes to the risk functions from each class are done in a way completely analogous to §5.1, and the colors in each plot correspond to results from a particular choice of risk function. The left-most plots show the regression lines obtained after the final iteration, whereas the remaining two plots show how the empirical average of the base and modified losses change over the course of the learning process. It is rather interesting to observe that between risk classes, we see very different base loss trajectories, but the end result in terms of the regression line learned is surprisingly uniform, although there are a few notable exceptions. One such exception is the entropic risk; for positive values of $\gamma$, the end result never varied far from the ordinary least-squares (OLS) solution, i.e., the traditional ERM solution under the quadratic base loss. In addition, we can see that changing $\alpha$ alone in M-risk and T-risk alone has virtually no impact on the long-term end result (we always end up at the OLS solution), but modifying $\sigma$ dramatically changes the evaluation. As a general trend, we see that changes to the risk made here essentially modulate between the OLS solution, and a solution that puts more weight on the extreme minority. It is interesting to note that this minority is not associated with large losses at initialization (reflected by the gray line), but its impact arises later in the learning process.

In Figure 12 we plot the results for the second round of experiments, where we fix individual risk functions and look at the impact of a modified base loss function, noting that colors now
correspond to the choice of loss function (i.e., the value of $1 \leq c \leq 2$). General trends for the trajectories are uniform across the risk classes, but we see a definite difference in terms of the degree to which the final result can be “pulled toward the majority,” in other words, the degree to which one can modulate between the least-squares solution and the least absolute error solution. We note that in particular CVaR does not move much, but additional experiments show that (as one would expect), the degree of shift due to changes in $c$ increases as we take $\beta$ closer to zero. Taking the results of Figure 12 together with those from the previous paragraph, we see clear evidence of how certain subsets of very different risk families can be used to achieve sensitivity to outliers (in terms of the final solution), despite having very different feedback trajectories leading to each solution.

5.4 Classification error trajectories

Our last angle of analysis will be to look at learning algorithms run on real-world classification datasets. We saw in §5.2 that even under a fixed model and normalized features, different datasets lead to very diverse loss distributions at initialization. Here we take the next step, and see what happens over the learning process for iterative algorithms fueled by distinct risk functions.

5.4.1 Experimental design

In essence, we run multi-class logistic regression on the well-known benchmark datasets seen previously in §5.2 (recall Figure 9), implemented using stochastic gradient descent (with averaging), and we record the base loss (logistic loss) and zero-one loss (misclassification error) at the end of each epoch. For perspective, we also record these losses at initialization, and store the full base loss distribution after the last update is finished. We run 10 independent trials, in which the data is shuffled and initial parameters are determined randomly. In each trial, for all datasets and risk classes, we use a mini-batch size of 32, and we run 30 epochs. As a reference algorithm, we run “vanilla” ERM, using the traditional risk $E_{\mu} L(h)$. For each alternative risk class (T-risk, CVaR, entropic, $\chi^2$-DRO), we select four specific risk functions, and for each, train using four possible step sizes (common across all risk classes).\footnote{We have also run tests on M-risk, but the results are quite similar to that of the T-risk here, so we have not included them here, instead using the space to show results for vanilla ERM as reference.} 80% of the data is used for training, 10% for validation, and 10% for testing. All the results we present are loss values computed on the test set. Validation data is used to evaluate different step sizes and choose the best one for each risk.

5.4.2 Results and discussion

In Figure 13, we give the loss trajectories for each of the risk families of interest, plus vanilla ERM, for three well-known benchmark datasets. Perhaps the most salient feature in these plots is that as the sensitivity to outlying data points is increased (e.g., larger $\alpha$, larger $\beta$, etc.), the trajectories deviate further from the vanilla ERM trajectory, both in terms of the base logistic loss and the classification error (zero-one loss). We remark here that while it may look like the $\chi^2$-DRO risk is rather insensitive to changes in $\tilde{a}$, it is only insensitive up to a certain threshold (just slightly beyond $\tilde{\eta} \approx 0.35$), after which performance rapidly deteriorates, making it quite difficult to modulate in a reliable manner across diverse datasets. The same statement holds for CVaR risk and entropic risk for larger values of $\beta$ and $\gamma$ than were tested here. On the other hand, T-risk over all valid choices of $\alpha$ (recalling Lemma 6), with $\sigma = 1.01$ and $\tilde{\eta} = 1.0$ fixed across all cases, does not change much at all. Modifications to $\sigma$ and $\tilde{\eta}$ were
not explored in these initial experiments, but as we have learned from §5.1, both play a key role in how T-risk behaves, and are natural avenues for subsequent analysis.

To complement the loss trajectories just discussed, in Figure 14 we give histograms of the base loss distribution on the test data for a particular trial (the first trial). We see that despite having error trajectories that can be brought quite close to that of vanilla ERM, and despite the fact that the learning process is very noisy, run very close to numerical convergence, the alternative risk classes are able to realize loss distributions quite distinct from those realized by ERM. While our understanding still remains quite imprecise, it is worth remarking how all of the alternative risk classes tend to yield loss distributions with stark differences in dispersion and symmetry when contrasted with vanilla ERM. This is an important fact, since the entire point of using alternative risks is to encourage the learner to realize a (test) loss distribution with certain desirable properties. Controlling spread and symmetry while keeping average performance strong is a natural application, and we see that the T-risk class is a safe bet across a wide variety of datasets and parameter settings. Further inquiry into the differences that arise in the loss distribution during the learning process for each risk class is an important future direction.
5.5 Figures for empirical analysis

Figure 3: An illustration of how each risk class depends on properties of the underlying distribution, with more details in the main text of §5.1. Reading left to right, in the first four plots, black curves denote risk values (dashed) and optimal thresholds (solid) as the risk function in each class is modified. The fifth plot is a histogram of the random sample \( \{L_1, \ldots, L_m\} \) used to estimate the risk, here from a centered Exponential distribution. Top row: original data. Bottom row: flipped data. Horizontal rules are drawn at the median (red, solid) and at zero (gray, dotted), and all plots share a common vertical axis.

Figure 4: Analogous to Figure 3, this time based on a centered Gamma distribution.
Figure 5: A more macroscopic look at the same risk classes seen in Figures 3–4, for a wider variety of underlying distribution families. We denote by \( q_b \) the \( b \)th empirical percentiles, and denote the empirical mean and median by blue and red horizontal rules, respectively.
Figure 6: Risk analysis specialized to the M-risk and T-risk classes, using the exact same distributions that were used in the previous experiments of Figure 5.
Figure 7: In the left-most plot, we plot the data (black scatter plot) and two linear binary classifier candidates (red-dashed and blue-dotted lines). The remaining three plots show histograms of the loss distributions incurred by each of these candidates using three loss functions common to classification.

Figure 8: Here the left-most plot is the (normalized) Belgian phone call dataset for use in a one-dimensional regression task, with two linear regression candidates. The remaining three plots show histograms of the loss distributions incurred by each of these candidates using three loss functions commonly used in regression.

Figure 9: Histogram of the (multi-class) logistic loss distribution incurred using a randomly initialized linear classifier, on both training (black) and test (gray) subsets of eight well-known datasets for classification.
Figure 10: Here we show the final regression lines learned under a variety of risk choices (left-most plot), and complement this with base and transformed average loss trajectories for each setting (two right-most plots). In this figure, consider just the M-risk and T-risk classes, under changes to $\alpha$ and $\sigma$. Gradient descent is initialized at the candidate shown in gray in the left-most plot.
Figure 11: Completely analogous to Figure 10, but looking at the CVaR, entropic, and $\chi^2$-DRO risk classes.
Figure 12: Analogous to Figures 10–11, except that now risk functions are fixed, and base losses are modified.
Figure 13: Average test losses (base and zero-one) as a function of epoch number, noting that epoch 0 is the initialization step. For each row, the vertical axis is common across all plots in the row. Plotted values are averages over all trials, and error bars denote ± standard deviation over trials.
Figure 14: Base loss distributions on the test data for the methods and datasets evaluated in the preceding Figure 13. Histograms are horizontal, with vertical axes common within each row.
Acknowledgments

This work was supported by JST ACT-X Grant Number JPMJAX200O and JST PRESTO Grant Number JPMJPR21C6.

References

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

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

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

[4] 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.

[5] Bassett, Jr, G. W. (1987). The St. Petersburg paradox and bounded utility. History of Political Economy, 19(4):517–523.

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

[7] Ben-Tal, A. and Teboulle, M. (1986). Expected utility, penalty functions, and duality in stochastic nonlinear programming. Management Science, 32(11):1445–1466.

[8] Ben-Tal, A. and Teboulle, M. (2007). An old-new concept of convex risk measures: The optimized certainty equivalent. Mathematical Finance, 17(3):449–476.

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

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

[11] Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. The Annals of Statistics, 43(6):2507–2536.

[12] Cardoso, A. R. and Xu, H. (2019). Risk-averse stochastic convex bandit. In 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), volume 89 of Proceedings of Machine Learning Research, pages 39–47.

[13] 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.

[14] Cutkosky, A. and Mehta, H. (2021). High-probability bounds for non-convex stochastic optimization with heavy tails. arXiv preprint arXiv:2106.14343v1.

[15] Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. SIAM Journal on Optimization, 29(1):207–239.
[16] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

[17] Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750v6*.

[18] Duchi, J. and Namkoong, H. (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55.

[19] Föllmer, H. and Knispel, T. (2011). Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 11(02n03):333–351.

[20] García, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

[21] Ghadimi, S., Lan, G., and Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305.

[22] Gotoh, J.-y., Kim, M. J., and Lim, A. E. (2018). Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters*, 46(4):448–452.

[23] Hacking, I. (2006). *The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2nd edition.

[24] 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.

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

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

[27] Holland, M. J. and Haress, E. M. (2021a). 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*.

[28] Holland, M. J. and Haress, E. M. (2021b). Spectral risk-based learning using unbounded losses. *arXiv preprint arXiv:2105.04816*.

[29] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, 2nd edition.

[30] Kashima, H. (2007). Risk-sensitive learning via minimization of empirical conditional value-at-risk. *IEICE Transactions on Information and Systems*, 90(12):2043–2052.

[31] Kearns, M. J. and Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press.
[32] 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.

[33] Koltchinskii, V. I. (1997). $m$-estimation, convexity and quantiles. The Annals of Statistics, pages 435–477.

[34] 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.

[35] Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2021a). On tilted losses in machine learning: Theory and applications. arXiv preprint arXiv:2109.06141v1.

[36] Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2021b). Tilted empirical risk minimization. In The 9th International Conference on Learning Representations (ICLR).

[37] Luenberger, D. G. (1969). Optimization by Vector Space Methods. John Wiley & Sons.

[38] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. Foundations of Computational Mathematics, 19(5):1145–1190.

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

[40] Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In Proceedings of the 22nd Conference on Learning Theory (COLT).

[41] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of Machine Learning. MIT Press.

[42] Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with $f$-divergences. In Advances in Neural Information Processing Systems 29 (NIPS 2016), volume 29, pages 2208–2216.

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

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

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

[46] Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. Journal of Banking & Finance, 26(7):1443–1471.

[47] Rockafellar, R. T., Uryasev, S., and Zabarankin, M. (2006). Generalized deviations in risk analysis. Finance and Stochastics, 10(1):51–74.

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

[49] Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.
A Detailed proofs

A.1 Proofs for section 2

A.1.1 Details for Barron class limits

For the limit as $\alpha \to 0$, use the fact that for any $a > 0$, we have

$$\lim_{x \to 0_+} \frac{(ax - 1)}{x} = \log(a).$$  \hfill (47)

This equality is sometimes known as Halley’s formula.\footnote{https://planetmath.org/HalleysFormula} For the limit as $\alpha \to -\infty$, first note that for any $\alpha < 0$ we can write $|2 - \alpha| = 2 + |\alpha|$, and thus $|\alpha|/2 = (2 - \alpha)/2 - 1$. With this in mind, for any $\alpha < 0$ we can observe

$$\left(1 + \frac{x^2}{|\alpha - 2|}\right)^{\alpha/2} = \left(1 + \frac{x^2}{|\alpha - 2|}\right)^{-|\alpha|/2} = \left(1 + \frac{x^2}{|\alpha - 2|}\right)^{1-(|2-\alpha|/2)} = \frac{\left(1 + \frac{x^2}{|\alpha - 2|}\right)^{|\alpha|/2}}{\sqrt{(1 + x^2)|\alpha - 2|}} = \frac{1}{\sqrt{\exp(x^2)}} = \exp(-x^2/2),$$

where the limit is taken as $\alpha \to -\infty$, and follows from the classical limit characterization of the exponential function. For the limit as $\alpha \to 2-$, first note that

$$|\alpha - 2| \left(1 + \frac{x^2}{|\alpha - 2|}\right)^{\alpha/2} = (|\alpha - 2|^{2/\alpha} + |\alpha - 2|^{(2/\alpha) - 1} x^2)^{\alpha/2}$$
and that as long as \( \alpha < 2 \), we can write
\[
|\alpha - 2|^{(2/\alpha) - 1} = (2 - \alpha)^{(2 - \alpha)/\alpha} = (u^u)^{1/(2 - u)}
\]
where we have introduced \( u := 2 - \alpha \). Taking \( \alpha \to 2_+ \) amounts to \( u \to 0_+ \), and thus using the fact that \( u^u \to 1 \) as \( u \to 0_+ \), the desired result follows from straightforward analysis.

### A.1.2 Details for entropic risk

Let \( X \sim \mu \) be an arbitrary random variable. Assuming the distribution is such that we can differentiate through the integral, we have
\[
\frac{d}{d\theta} \left[ \theta + \frac{1}{\gamma} \left( E\mu e^{\gamma(X-\theta)} - 1 \right) \right] = 0 \iff E\mu e^{\gamma(X-\theta)} = 1.
\]

Let \( \theta^* \) be any value that satisfies the first-order optimality condition in (48). It follows that
\[
\theta^* + \frac{1}{\gamma} \left( E\mu e^{\gamma(X-\theta^*)} - 1 \right) = \theta^*.
\]

It is easy to confirm that setting \( \theta^* = (1/\gamma) \log(E\mu e^{\gamma X}) \) gives us a valid solution.

### A.2 Proofs for section 3

**Proof of Lemma 2.** In this proof, without further mention, we will make regular use of the following two helper results: Lemma 13 (bounded gradient implies Lipschitz continuity) and Lemma 14 (positive definite Hessian implies convexity). For reference, the first and second derivatives of \( \rho_\sigma \) are given in §B.3. We take up each \( \alpha \) setting one at a time.

First, the case of \( \alpha = 2 \). For this case, clearly \( \rho_\sigma' \) is unbounded, and thus \( \rho_\sigma \) is not (globally) Lipschitz on \( \mathbb{R} \). On the other hand, since \( \rho_\sigma''(x) = 1/\sigma^2 \), we have that \( \rho_\sigma' \) is \( \lambda \)-Lipschitz with \( \lambda = (1/\sigma^2) \).

Next, the case of \( \alpha = 0 \). For any fixed \( \sigma > 0 \), in both the limits \( x \to 0 \) and \( |x| \to \infty \), we have \( \rho_\sigma'(x) \to 0 \). Maximum and minimum values are achieved when \( \rho_\sigma'(x) = 0 \), and this occurs if and only if \( x^2 = 2\sigma^2 \). It follows from direct computation that \( \rho_\sigma'(\pm \sqrt{2}\sigma) = \pm 1/(\sqrt{2}\sigma) \), and thus \( \rho_\sigma \) is \( \lambda \)-Lipschitz with \( \lambda = 1/(\sqrt{2}\sigma) \). Next, recalling that \( \rho_\sigma'' \) takes the form
\[
\rho_\sigma''(x) = \frac{2}{x^2 + 2\sigma^2} \left( 1 - \frac{2x^2}{x^2 + 2\sigma^2} \right),
\]
we see that this is a product of two factors, one taking values in \((0, 1/\sigma^2]\), and one taking values in \((-1, 1]\). The absolute value of both of these factors is maximized when \( x = 0 \), and so \( |\rho_\sigma''(x)| \leq |\rho_\sigma''(0)| = 1/\sigma^2 \), meaning that \( \rho_\sigma' \) is \( \lambda \)-Lipschitz with \( \lambda = 1/\sigma^2 \). Finally, regarding convexity, we have that \( \rho_\sigma''(x) \geq 0 \) if and only if \( |x| \leq \sqrt{2}\sigma \).

Next, the case of \( \alpha = -\infty \). For any fixed \( \sigma > 0 \), we have \( \rho_\sigma'(x) \to 0 \) in both the limits \( x \to 0 \) and \( |x| \to \infty \). Furthermore, it is immediate that \( \rho_\sigma''(x) = 0 \) at the points \( x = \pm \sigma \). Evaluating \( \rho_\sigma' \) at these stationary points we have \( \rho_\sigma'(\pm \sigma) = \pm (1/\sigma) \exp(-1/2) \), and thus \( \rho_\sigma \) is \( \lambda \)-Lipschitz with \( \lambda = (1/\sigma) \exp(-1/2) \). Regarding bounds on \( \rho_\sigma'' \), first note that \( \rho_\sigma''(x) \to 0 \) as \( |x| \to \infty \), and \( \rho_\sigma''(0) = 1/\sigma^2 \). Then to identify stationary points, note that
\[
\rho_\sigma'''(x) = \frac{1}{\sigma^3} \exp \left( -\frac{1}{2} \frac{x}{\sigma} \right) \left( \frac{x^2}{\sigma^2} - 1 \right) - \frac{2x}{\sigma^3}
\]

38
and thus $\rho''_\sigma(x) = 0$ if and only if $(x/\sigma)^2 - 1 = 2$, i.e., the stationary points are $x = \pm \sqrt{3}\sigma$, both of which yield the same value, namely $\rho''_\sigma(\pm \sqrt{3}\sigma) = -(2/\sigma^2) \exp(-3/2)$. Since $2 \exp(-3/2) \approx 0.45 < 1$, we conclude that $\rho''_\sigma$ is $\lambda$-Lipschitz with $\lambda = 1/\sigma^2$. Finally, since $\rho''_\sigma(x) \geq 0$ if and only if $|x| \leq \sigma$, this specifies the region on which $\rho_\sigma$ is convex.

Finally, all that remains is the general case of $-\infty < \alpha < 2$ where $\alpha \neq 0$. Note that in order for $\rho'_\sigma(x) = 0$ to hold, we require

$$\frac{2(x/\sigma)^2}{1 + (x/\sigma)^2 / |\alpha - 2|} = \frac{|\alpha - 2|}{1 - (\alpha/2)},$$

which via some basic algebra is equivalent to

$$\left(\frac{x}{\sigma}\right)^2 = \frac{|\alpha - 2|}{1 - \alpha}.$$

Clearly, this is only possible when $\alpha < 1$, so we consider this sub-case first. This implies stationary points $\pm x^* := \pm \sigma \sqrt{|\alpha - 2|/(1 - \alpha)}$, for which we have

$$\rho'_\sigma(\pm x^*) = \pm \frac{1}{\sigma} \sqrt{\frac{|\alpha - 2|}{1 - \alpha}} \left(\frac{2 - \alpha}{1 - \alpha}\right)^{(\alpha/2)-1} = \pm \frac{1}{\sigma} \left(\sqrt{\frac{|\alpha - 2|}{1 - \alpha}}\right)^{\alpha-1} = \pm \frac{1}{\sigma} \left(\frac{1}{\sqrt{|\alpha - 2|}}\right)^{1-\alpha}.$$

Since $\rho'_\sigma(x) \rightarrow 0$ in both the limits $x \rightarrow 0$ and $|x| \rightarrow \infty$, we have obtained a maximum value for $\rho'_\sigma$ at $x^*$, thus implying for the case of $\alpha < 1$ that $\rho_\sigma$ is $\lambda$-Lipschitz, with a coefficient of $\lambda = (1/\sigma)(\sqrt{1 - \alpha}/|\alpha - 2|)^1\alpha$. For the case of $\alpha = 1$, direct inspection shows

$$|\rho'_\sigma(x)| = \frac{|x/\sigma^2|}{1 + (x/\sigma)^2} = \frac{1}{\sigma^2 \sqrt{(1/x^2) + (1/\sigma^2)}},$$

a value which is maximized in the limit $|x| \rightarrow \infty$. As such, for $\alpha = 1$, we have that $\rho_\sigma$ is $\lambda$-Lipschitz with $\lambda = 1/\sigma$. For the case of $1 < \alpha < 2$, $\rho'_\sigma$ is unbounded. To see this, note that for $x > 0$ we have

$$\rho'_\sigma(x) = \frac{1}{\sigma^2} \left(\frac{1 + (x/\sigma)^2 / |\alpha - 2|}{(1/x) + x/(\sigma^2 |\alpha - 2|)}\right) \left(\frac{1}{\sigma^2 + \alpha \sqrt{|\alpha - 2|}}\right) \text{exp}\left(\frac{x^\alpha}{(1/x) + x/(\sigma^2 |\alpha - 2|)}\right),$$

and since $\alpha > 1$, sending $x \rightarrow \infty$ clearly implies $\rho'_\sigma(x) \rightarrow \infty$, and this means that $\rho_\sigma$ cannot be Lipschitz on $\mathbb{R}$ when $\alpha > 1$. As for bounds on $\rho''_\sigma$, recall that

$$\rho''_\sigma(x) = \frac{1}{\sigma^2} \left(\frac{1 + (x/\sigma)^2 / |\alpha - 2|}{A(x)}\right)^{(\alpha/2)-1} \left(1 - \frac{1 - (\alpha/2)}{|\alpha - 2|} \frac{2(x/\sigma)^2}{1 + (x/\sigma)^2 / |\alpha - 2|} \right),$$

where we have introduced the labels $A(x)$ and $B(x)$ just as convenient notation. Fixing any $\sigma > 0$, first note that since $\alpha < 2$, we have $(\alpha/2) - 1 < 0$ and thus $0 \leq A(x) \leq 1$. Next, direct inspection shows $0 \leq B(x) \leq 2(1 - (\alpha/2))$. These two facts immediately imply an upper bound $\rho''_\sigma(x) \leq 1/\sigma^2$ and a lower bound $\rho''_\sigma(x) \geq -(1 - \alpha)/\sigma^2$, both of which hold for any $\alpha < 2$. Furthermore, for the case of $1 \leq \alpha < 2$, we thus have $0 \leq \rho''_\sigma(x) \leq 1/\sigma^2$. When $\alpha < 1$ however, $\rho''_\sigma$ can be negative. To get matching lower bounds requires $A(x)(1 - B(x)) \geq -1$, or

39
A(x)(B(x) - 1) \leq 1. To study conditions under which this holds, first note that B(x) can be re-written as
\[ B(x) = \left( \frac{2 - \alpha}{|\alpha - 2|} \right) \frac{(x/\sigma)^2}{1 + (x/\sigma)^2} = \frac{(x/\alpha)^2}{1 + (x/\alpha)^2}, \]
and thus we have
\[ A(x)(B(x) - 1) = \frac{(x/\alpha)^2}{\left(1 + (x/\alpha)^2\right)^{2-(\alpha/2)}} - \frac{1}{\left(1 + (x/\alpha)^2\right)^{1-(\alpha/2)}}. \tag{49} \]
To get a more convenient upper bound on this, observe that \((1 + x)^{1-(\alpha/2)} \leq (1 + x)^{2-(\alpha/2)}\) for any \(x \geq 0\) and \(-\infty \leq \alpha \leq 2\). It follows immediately that
\[ A(x)(B(x) - 1) \leq \frac{(x/\alpha)^2 - 1}{\left(1 + (x/\alpha)^2\right)^{2-(\alpha/2)}}. \tag{50} \]
To get the right-hand side of (50) to be no greater than 1 is equivalent to
\[ (x/\alpha)^2 - 1 \leq \left(1 + \frac{(x/\alpha)^2}{|\alpha - 2|}\right)^{2-(\alpha/2)}. \tag{51} \]
For the case of \(0 \leq \alpha < 1\), note that \(1 \leq |\alpha - 2| = 2 - \alpha < 2 - (\alpha/2)\), and using the helper inequality (72), we have
\[ \left(1 + \frac{(x/\alpha)^2}{|\alpha - 2|}\right)^{2-(\alpha/2)} \geq \left(1 + \frac{(x/\alpha)^2}{|\alpha - 2|}\right)^{|\alpha - 2|} \geq 1 + (x/\sigma)^2 > (x/\sigma)^2 - 1, \]
which implies (51) for \(0 \leq \alpha < 1\). All that remains is the case of \(-\infty < \alpha < 0\), which requires a bit more care. Returning to the exact form of \(A(x)(B(x) - 1)\) given in (49), note that the inequality
\[ (x/\sigma)^2 - \left(1 + \frac{(x/\sigma)^2}{|\alpha - 2|}\right) \leq \left(1 + \frac{(x/\sigma)^2}{|\alpha - 2|}\right)^{2-(\alpha/2)} \tag{52} \]
is equivalent to the desired property, i.e., \(A(x)(B(x) - 1) \leq 1 \iff (52)\). Using Bernoulli’s inequality (74), we can bound the right-hand side of (52) as
\[ \left(1 + \frac{(x/\sigma)^2}{|\alpha - 2|}\right)^{2-(\alpha/2)} \geq 1 + \left(\frac{2 - (\alpha/2)}{|\alpha - 2|}\right)(x/\sigma)^2. \]
Subtracting the left-hand side of (52) from the right-hand side of the preceding inequality, we obtain
\[ \left(1 + \frac{(x/\sigma)^2}{|\alpha - 2|}\right)^{2-(\alpha/2)} - [x/\sigma)^2 - 1 - \frac{(x/\sigma)^2}{|\alpha - 2|} \geq 2 + \left(\frac{2 - (\alpha/2)}{|\alpha - 2|} - 1 + \frac{1}{|\alpha - 2|}\right)(x/\sigma)^2 \]
\[ = 2 + \left(\frac{1 - (\alpha/2)}{2 + |\alpha|}\right)(x/\sigma)^2, \tag{53} \]
where the second step uses the fact that for \(\alpha < 0\), we can write \(|\alpha - 2| = 2 + |\alpha|\) and \(2 - (\alpha/2) = 2 + (|\alpha|/2)\). Note that the right-hand side of (53) is non-negative for all \(x \in \mathbb{R}\).
whenever \(-2 \leq \alpha < 0\), which via (52) tells us that \(A(x)(B(x) - 1) \leq 1\) indeed holds in this case as well. For the case of \(-\infty < \alpha < -2\), note that showing (52) holds is equivalent to showing \(f_\alpha(x) \geq 0\) for all \(x \geq 0\), where for convenience we define the polynomial

\[
f_\alpha(x) := 1 + \left(\frac{1}{2 + |\alpha|} - 1\right)x + \left(1 + \frac{x}{2 + |\alpha|}\right)^2 + (|\alpha|/2).
\]

The first derivative is

\[
f'_\alpha(x) = 2 + (|\alpha|/2)\left(1 + \frac{x}{2 + |\alpha|}\right) + \frac{1}{2 + |\alpha|} - 1,
\]

and with this form in hand, solving for \(f_\alpha(x) = 0\), it is straightforward to confirm that \(x^*_\alpha\), given below is a stationary point:

\[
x^*_\alpha := (2 + |\alpha|)\left[\left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)} - 1\right].
\]

Furthermore, it is clear that \(f''_\alpha \geq 0\), implying that \(f_\alpha\) is convex, and that \(x^*_\alpha\) is a minimum. As such, the minimum value taken by \(f_\alpha\) on \(\mathbb{R}_+\) is

\[
f_\alpha(x^*_\alpha) = \left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)} - (2 + |\alpha|)\left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)} - \frac{1}{2 + (|\alpha|/2)}\left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{2/(|\alpha|/2)}
\]

\[
= (2 + |\alpha|) + \left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{2/(|\alpha|/2)} - (1 + |\alpha|)\left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)}\left[1 - \frac{1}{2 + (|\alpha|/2)}\right]
\]

\[
= 1 + (1 + |\alpha|)\left[1 - \left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)}\left[1 - \frac{1}{2 + (|\alpha|/2)}\right]\right].
\]

We require \(f_\alpha(x^*_\alpha) \geq 0\) for all \(-\infty < \alpha < -2\). From the preceding equalities, note that a simple sufficient condition for \(f_\alpha(x^*_\alpha) \geq 1\) is

\[
\left(\frac{1 + |\alpha|}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)}\left[1 - \frac{1}{2 + (|\alpha|/2)}\right] \leq 1
\]

or equivalently

\[
\left(1 - \frac{1}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)} \leq \frac{2 + (|\alpha|/2)}{1 + |\alpha|}.
\]

Applying the helper inequality (73) to the left-hand side of (54), we have

\[
\left(1 - \frac{1}{2 + (|\alpha|/2)}\right)^{1/(|\alpha|/2)} \leq 1 + \frac{\left(-1 + (|\alpha|/2)\right)}{\left(1 - \frac{1}{2 + (|\alpha|/2)}\right)} = 1 - \frac{1 + (|\alpha|/2)}{2 + |\alpha|} = \frac{1 + (|\alpha|/2)}{2 + |\alpha|}
\]

\[
\leq \frac{2 + (|\alpha|/2)}{1 + |\alpha|}.
\]

This is precisely the desired inequality (54), implying \(f_\alpha(x^*_\alpha) \geq 1 > 0\) for all \(-\infty < \alpha < -2\), and in fact all real \(\alpha < 0\). To summarize, we have \(A(x)(B(x) - 1) \leq 1\) for all \(x \in \mathbb{R}\), and thus the desired \(1/\sigma^2\)-smoothness result follows, concluding the proof. 

41
Proof of Lemma 3. Let $X$ denote any $\mathcal{F}$-measurable random variable. The continuity of $\rho$ implies that the integral $E_\mu \rho_\sigma (X - \theta)$ exists; we just need to prove it is finite.\footnote{This uses the fact that any composition of (Borel) measurable functions is itself measurable \cite[Lem. 1.5.7]{1}.} Since we are taking $\rho$ from the Barron class (5), we consider each $\alpha$ setting separately. Starting with $\alpha = 2$, note that

$$\rho_\sigma (X - \theta; 2) = \frac{1}{2} \left( \frac{X - \theta}{\sigma} \right)^2$$

and thus $E_\mu X^2 < \infty$ is sufficient and necessary. For $\alpha = 0$, first note that we have

$$\rho_\sigma (X - \theta; 0) = \log \left( 1 + \frac{1}{2} \left( \frac{X - \theta}{\sigma} \right)^2 \right).$$

Let $f_1(x) \equiv \log(1 + x)$ and $f_2(x) \equiv x^c/c$, where $0 < c < 1$. Note that $f_1(0) = f_2(0) = 0$, and furthermore that for any $x > 0$,

$$f_1'(x) = \frac{1}{1 + x} < \left( \frac{1}{1 + x} \right)^{1-c} < \left( \frac{1}{x} \right)^{1-c} = f_2'(x).$$

We may thus conclude that $f_1(x) \leq f_2(x)$ for all $x \geq 0$, and thus for any $0 < c < 1$ we have

$$\log \left( 1 + \frac{1}{2} \left( \frac{X - \theta}{\sigma} \right)^2 \right) \leq \frac{1}{c} \left( \frac{X - \theta}{\sqrt{2}\sigma} \right)^{2c}.$$

It follows that to ensure $E_\mu \rho_\sigma (X - \theta; 0) < \infty$, it is sufficient if we assume $E_\mu |X|^c < \infty$ for some $c > 0$. Proceeding to the case of $\alpha = -\infty$, we have

$$\rho_\sigma (X - \theta; -\infty) = 1 - \exp \left( -\frac{1}{c} \left( \frac{X - \theta}{\sigma} \right)^2 \right).$$

Any composition of measurable functions is measurable, and since the right-hand side is bounded above by 1 and below by 0, we have that $\rho_\sigma (X - \theta; -\infty)$ is $\mu$-integrable without requiring any extra assumptions on $X$ besides measurability. All that remains for the Barron class is the case of non-zero $-\infty < \alpha < 2$, where we have

$$\rho_\sigma (X - \theta; \alpha) = \frac{\alpha - 2}{\alpha} \left( 1 + \frac{1}{\alpha - 2} \left( \frac{X - \theta}{\sigma} \right)^2 \right)^{\alpha/2} - 1.$$

Let us break this into two cases: $-\infty < \alpha < 0$ and $0 < \alpha < 2$. Starting with the former case, this is easy since

$$\left( 1 + x^2 \right)^{\alpha/2} = \frac{1}{\left( \sqrt{1 + x^2} \right)^{-\alpha}}$$

which is bounded above by 1 and below by 0 for any $\alpha < 0$ and $x \in \mathbb{R}$, which means the random variable $\rho_\sigma (X - \theta; \alpha)$ is $\mu$-integrable without any extra assumptions on $X$. As for the latter case of $0 < \alpha < 2$, first note that the monotonicity of $(\cdot)^{\alpha/2}$ on $\mathbb{R}_+$ implies

$$(1 + x^2)^{\alpha/2} \geq |x|^\alpha$$

which means $E_\mu |X|^\alpha < \infty$ is necessary. That this condition is also sufficient is immediate from the form of $\rho_\sigma (X - \theta; \alpha)$ just given. This concludes the proof; the desired result stated in the lemma follows from setting $X = L(h)$ and the observation that the choice of $\theta \in \mathbb{R}$ in the preceding discussion was arbitrary.
Proof of Lemma 5. Referring to the derivatives (70)–(71) in §B.3, we know that \( \rho'_\alpha \) is measurable, and by the proof of Lemma 2, we know that \( \| \rho' \|_\infty < \infty \) for all \( \alpha \leq 1 \). Thus, as long as \( \text{L}(h) \) is \( \mathcal{F} \)-measurable, we have that \( \rho'_\alpha(\text{L}(h) - \theta) \) is \( \mu \)-integrable. For the case of \( 1 < \alpha \leq 2 \), note that \( |\rho'_\alpha(x)| \leq |x|/\sigma^2 \) holds, meaning that \( \mathbb{E}_\mu[\text{L}(h)] < \infty \) implies integrability. Similarly for the second derivatives, from the proof of Lemma 2, we see that \( \| \rho'' \|_\infty < \infty \) for all \( -\infty \leq \alpha \leq 1 \), implying the \( \mu \)-integrability of \( \rho''_\alpha(\text{L}(h) - \theta) \).

The Leibniz integration property follows using a straightforward dominated convergence argument, which we give here for completeness. Letting \( (a_k) \) be any non-zero real sequence such that \( a_k \to 0 \), we can write

\[
\left. \frac{d}{d\theta} \text{D}_\rho(h; \theta) = \lim_{k \to \infty} \frac{\text{D}_\rho(h; \theta + a_k) - \text{D}_\rho(h; \theta)}{a_k} \right|_{a_k}
\]

For notational convenience, let us denote the key sequence of functions by

\[
f_k := \rho_\alpha(\text{L}(h) - (\theta + a_k)) - \rho_\alpha(\text{L}(h) - \theta)
\]

and note that \( f_k \to f := -\rho'_\alpha(\text{L}(h) - \theta) \) pointwise as \( k \to \infty \). We can then say the following: for all \( k \), we have that

\[
|f_k| \leq \sup_{0 < c < 1} |\rho'_\alpha(\text{L}(h) - (\theta + ca_k))| \leq g := |\rho'_\alpha(\text{L}(h) - \theta')|
\]

for an appropriate choice of \( \theta' \in \mathbb{R} \). The first inequality follows from the helper Lemma 13. We can always find an appropriate \( \theta' \) because the sequence \( (a_k) \) is bounded and \( \rho' \) is eventually monotone, regardless of the choice of \( \alpha \). With the fact \( |f_k| \leq g \) in hand, recall that we have already proved that \( \mathbb{E}_\mu g < \infty \) under the assumptions we have made, and thus \( \mathbb{E}_\mu f_k \to \mathbb{E}_\mu f \) by dominated convergence.\(^\text{27}\) As such, we have

\[
\left. \frac{d}{d\theta} \text{D}_\rho(h; \theta) = \lim_{k \to \infty} \mathbb{E}_\mu f_k = \mathbb{E}_\mu f = -\mathbb{E}_\mu \rho'_\alpha(\text{L}(h) - \theta), \right|_{\mathbb{E}_\mu}
\]

which is the desired Leibniz property for the first derivative. A completely analogous argument holds for the second derivative, yielding the desired result. \( \square \)

\(^{\text{27}}\) See for example Ash and Doléans-Dade [3, Thm. 1.6.9].
Next, considering the threshold risk $\tilde{R}_\rho$, since we are doing unconstrained optimization, any solution $\tilde{\theta}^*$ must satisfy the first-order condition $\tilde{h} + d_\theta D_\rho(h; \tilde{\theta}^*) = 0$, where $d_\theta := d/d\theta$. Using Lemma 5 again, this can be equivalently re-written as

$$
E_\mu \rho'_\sigma \left(L(h) - \tilde{\theta}^*\right) = \tilde{h}.
$$

(55)

When $\alpha > 1$, the derivative of the dispersion function has unbounded range, i.e., $\rho'_\sigma(\mathbb{R}) = \mathbb{R}$. As such, an argument identical to that used in the proof of Lemma 15 implies that for any $\tilde{\eta} \in \mathbb{R}$, we can always find a $\tilde{\theta}^* \in \mathbb{R}$ such that (55) holds exactly, recalling that we have continuity via Lemma 5. Combining this with convexity gives us a valid solution. The special case of $\alpha = 1$ requires additional conditions, since from Lemma 2, we know that in this case $|\rho'_\sigma| \leq 1/\sigma$, and thus by an analogous argument, whenever $|\tilde{\theta}^*| < 1/\sigma$ we can find a finite solution.

To prove uniqueness in the special case of $1 \leq \alpha \leq 2$, direct inspection of the second derivative in (71) shows us that $\rho''_\sigma(x) > 0$ on $\mathbb{R}$ whenever we have

$$
\frac{2(x/\sigma)^2}{1 + (x/\sigma)^2} < \frac{|\alpha - 2|}{1 - (\alpha/2)}.
$$

Re-arranging the above inequality yields an equivalent condition of $(x/\sigma)^2(1 - \alpha) < |\alpha - 2|$, a condition which holds on $\mathbb{R}$ if and only if $1 \leq \alpha \leq 2$. Since $\rho'_\sigma$ is positive on $\mathbb{R}$, this implies that $E_\mu \rho''_\sigma(L(h) - \theta; \alpha) > 0$ for all $\theta \in \mathbb{R}$. Using Lemma 5, we have that $E_\mu \rho''_\sigma(L(h) - \theta; \alpha)$ is equal to the second derivative of $D_\rho(h; \theta)$ with respect to $\theta$, which implies that $\theta \mapsto D_\rho(h; \theta)$ and $\theta \mapsto D_\rho(h; \theta) + \tilde{\eta} \theta$ are strictly convex on $\mathbb{R}$, and thus their minimum must be unique. 



Proof of Lemma 7. For random loss $L$, using Lemma 5, first-order optimality conditions require

$$
E_\mu \rho'_\sigma \left(L - \theta^*_L\right) = 0, \quad E_\mu \rho'_\sigma \left(L - \tilde{\theta}^*_L\right) = \tilde{h}.
$$

(56)

If these conditions hold, then from direct inspection, the same conditions will clearly hold if we replace $L$ by $L + c$, $\theta^*_L$ by $\theta^*_L + c$, and $\tilde{\theta}^*_L$ by $\tilde{\theta}^*_L + c$. This implies both translation-invariance of the dispersions and the translation-equivalence of the optimal thresholds. Non-negativity follows trivially from the fact that $\rho(\cdot) \geq 0$. Noting that $\rho(x) > 0$ for all $x \neq 0$, we have that $D_\rho(L; \theta) = 0$ if and only if $L = \theta$ almost surely. Since $D_\rho(L; \theta^*_L) \leq D_\rho(L; \tilde{\theta}^*_L)$ by the optimality of $\theta^*_L$, it follows that for any non-constant $L$, we must have $D_\rho(L; \tilde{\theta}^*_L) > 0$. Furthermore, from the optimality condition (56) for $\tilde{\theta}^*_L$, even when $L$ is constant, we must have $D_\rho(L; \tilde{\theta}^*_L) > 0$ whenever $\tilde{h} \neq 0$, since $\rho'_\sigma(x) = 0$ if and only if $x = 0$.

In the special case where $1 \leq \alpha \leq 2$, we have that $\rho''_\sigma$ is positive on $\mathbb{R}$ (see §B.3 and Fig. 1). This implies that $\rho_\sigma$ is strictly convex, and $\rho'_\sigma$ is monotonically increasing. Let $L_1 \leq L_2$ almost surely, but say $\theta^*_L > \tilde{\theta}^*_L$. Using the optimality condition (56), uniqueness of the solution via Lemma 6, and the aforementioned monotonicity of $\rho'_\sigma$, we have

$$
0 = E_\mu \rho'_\sigma (L_1 - \theta^*_L) < E_\mu \rho'_\sigma (L_1 - \tilde{\theta}^*_L) \leq E_\mu \rho'_\sigma (L_2 - \tilde{\theta}^*_L) = 0.
$$

\[28\] See for example Boyd and Vandenberge [10, Sec. 3.1.4].

\[29\] This fact follows from basic Lebesgue integration theory [3, Thm. 1.6.6].
This is a contradiction, and thus we must have \( \theta^*_{L_1} \leq \theta^*_{L_2} \). An identical argument using the exact same properties proves that \( \tilde{\theta}^*_{L_1} \leq \tilde{\theta}^*_{L_2} \) also holds. Finally, to prove convexity, take any \( L_1, L_2 \in L, \theta_1, \theta_2 \in \mathbb{R}, \) and \( a \in (0, 1) \), and note that

\[
\tilde{\rho}(a L_1 + (1 - a) L_2) \leq D_{\rho}(a L_1 + (1 - a) L_2; a \theta_1 + (1 - a) \theta_2) + \eta(a \theta_1 + (1 - a) \theta_2)
\]

\[
= \mathbb{E}_\mu \rho_\sigma(a(L_1 - \theta_1) + (1 - a)(L_2 - \theta_2)) + \eta(a \theta_1 + (1 - a) \theta_2)
\]

\[
\leq a (D_{\rho}(L_1; \theta_1) + \eta \theta_1) + (1 - a) (D_{\rho}(L_2; \theta_2) + \eta \theta_2).
\]

The first inequality uses optimality of the threshold in the definition of \( \tilde{\rho}, \) whereas the second inequality uses the convexity of \( \rho_\sigma. \) Since the choice of \( \theta_1 \) and \( \theta_2 \) here were arbitrary, we can set \( \tilde{\theta}_1 = \tilde{\theta}^*_{L_1} \) and \( \tilde{\theta}_2 = \tilde{\theta}^*_{L_2} \) to obtain the desired inequality

\[
\tilde{\rho}(a L_1 + (1 - a) L_2) \leq a \tilde{\rho}(L_1) + (1 - a) \tilde{\rho}(L_2)
\]

giving us convexity of the threshold risk. As a direct corollary, setting \( \eta = 0 \) yields the convexity result for \( L \rightarrow D_{\rho}(L; \tilde{\theta}^*_L) \), which in turn implies the result for \( \rho, \) since the sum of convex and affine functions is convex. \( \square \)

### A.3 Proofs for section 4

**Proof of Lemma 8.** The crux of this result is an analogue to Lemma 5 regarding the differentials of \( D_\rho(h; \theta) \), this time taken with respect to \( h \), rather than \( \theta \). Fixing arbitrary \( g, h \in H \), let us start by considering the following sequence of random variables:

\[
f_k := \frac{\rho_\sigma(L(h + a_k g) - \theta) - \rho_\sigma(L(h) - \theta)}{a_k}
\]

where \( (a_k) \) is any sequence of real values such that \( a_k \rightarrow 0^+ \) as \( k \rightarrow \infty \). Before getting into the details, let us unpack the differentiability assumption made on the base loss. Before random sampling, the map \( h \mapsto L(h) \) is of course a map from \( H \) to the set of measurable functions \( \{L(h) : h \in H\}, \) but after sampling, there is no randomness and it is simply a map from \( H \) to \( \mathbb{R}. \) Having sampled the random loss, the property we desire is that for each \( h \in H, \) there exists a continuous linear functional \( L'(h) : U \rightarrow \mathbb{R} \) such that

\[
\lim_{||g|| \rightarrow 0} \frac{|L(h + g) - L(h) - L'(h)(g)|}{||g||} = 0.
\]

The differentiability condition in the lemma statement is simply that

\[
P\{ \text{equality (58) holds} \} = 1.
\]

On this “good” event, since the map \( x \mapsto \rho_\sigma(x) \) is differentiable by definition, we have that the composition \( h \mapsto \rho_\sigma(L(h) - \theta) \) is also differentiable for any choice of \( \theta \in \mathbb{R}, \) and a general chain rule can be applied to compute the differentials.\(^{30}\) In particular, we have a pointwise limit of

\[
f := \lim_{k \rightarrow \infty} f_k = \rho_\sigma'(L(h) - \theta)L'(h)(g)
\]

which also uses the fact that the Fréchet and Gateaux differentials are equal here.\(^{31}\) Technically, it just remains to obtain conditions which imply \( \mathbb{E}_\mu f_k \rightarrow \mathbb{E}_\mu f. \) In pursuit of a \( \mu \)-integrable

\(^{30}\)See Penot [44, Thm. 2.47] for this key fact, where “X” is \( U \) here, and both “Y” and “Z” are \( \mathbb{R} \) here.

\(^{31}\)Luenberger [37, §7.2, Prop. 2]
upper bound on the sequence \((f_k)\), note that for large enough \(k\), we have
\[
|f_k| \leq \frac{1}{a_k} \|a_k g\| \sup_{0 < a < a_k} \|\rho'_\sigma(L(h + ag) - \theta)L'(h + ag)\|
\]
\[
\leq \|g\| \sup \{\|\rho'_\sigma(L(h_0) - \theta)L'(h_0)\| : h_0 \in \mathcal{H}\}
\]
\[
\leq \frac{|g|}{\sigma^2} \sup \{\|L(h_0) - \theta\||L'(h_0)\| : h_0 \in \mathcal{H}\}. \tag{61}
\]

The key to the first of the preceding inequalities is a generalized mean value theorem.\(^{32}\) Both the first and second inequalities also use the fact that \(h + a_k g \in \mathcal{H}\) eventually. The final inequality uses the fact that \(\rho'_\sigma(x) = \rho'(x/\sigma)/\sigma \leq |x|/\sigma^2\) for any choice of \(-\infty \leq \alpha \leq 2\). This inequality suggests a natural condition of
\[
E_\mu \left[ \sup_{h_0 \in \mathcal{H}} \|L(h_0)L'(h_0)\| \right] < \infty \tag{62}
\]
under which we can apply a standard dominated convergence argument.\(^{33}\) In particular, the key implication is that
\[
(62) \implies \lim_{k \to \infty} E_\mu f_k = E_\mu f. \tag{63}
\]

Since we have
\[
\lim_{k \to \infty} E_\mu f_k = \lim_{a \to 0^+} E_\mu \left[ \frac{\rho'_\sigma(L(h + ag) - \theta) - \rho'_\sigma(L(h) - \theta)}{a} \right] = D'_\rho(h; \theta)(g),
\]
where \(D'_\rho(h; \theta) : \mathcal{U} \to \mathbb{R}\) denotes the gradient of \(h \mapsto D_\rho(h; \theta)\), we see that by applying the preceding argument (culminating in \((63)\)) to the modified losses (36) and (37), we readily obtain the desired result. \(\square\)

**Proof of Theorem 11.** As mentioned in \(\S 4.6\), with all the preparatory definitions and helper lemmas in place, what remains of proving this theorem is an almost entirely mechanical process based on well-established techniques; we refer to Ghadimi et al. \([21, \text{Cor. 3}]\) as a standard reference. Essentially, we just need to show that the properties we have established for our objectives \((r_\mu\) and \(\tilde{r}_\rho\)) and our additional technical assumptions can be aligned with the setup required by \([21]\). For completeness, we spell out the correspondence precisely here.

The quantity denoted by “\(\alpha\)” in Ghadimi et al. \([21]\) is set to 1 here, since we consider the special case of a quadratic distance generating function (their “\(\omega(\cdot)\)” which is 1-strongly convex. Our objectives \(r_\mu\) and \(\tilde{r}_\rho\) correspond to their “\(f\)” and the critical smoothness property follows from our Lemma 9. Their set “\(X\)” corresponds to our \(\mathcal{H} \times \mathbb{R}\); both are assumed to be closed and convex (via \(B_1\) here). The key quantity \(\Delta_T\) with random \(T\) based on \((46)\) corresponds to “\(g_{X,H}\)” in their notation. Their results are stated in terms of usual Euclidean space, but everything trivially extends to arbitrary Hilbert spaces, where the norm is induced by the underlying inner product. Our assumption \(B_2\) corresponds to their assumption “\(A_1(\alpha)\).” Finally, the variance factor in our upper bound corresponds to their “\(\sigma^2\)” bound. This variance is clearly finite and bounded above by the right-hand side of the inequality in \(A_1\). Taking these facts together, the assumptions needed to apply Corollary 3 of Ghadimi et al. \([21]\) are satisfied, implying the desired result. \(\square\)

\(^{32}\)Considering the proof of Lemma 13 due to Luenberger [37, \S 7.3, Prop. 2], just generalize the one-dimensional part of the argument from the original interval \([0,1]\) to the interval \([0,a_k]\) here.

\(^{33}\)See for example Ash and Doléans-Dade [3, Thm. 1.6.9]. If \((61)\) holds for say all \(k \geq k_0\), then we can just bound \(|f_k|\) by the greater of \(\max_{j \leq k_0} |f_j|\) (clearly \(\mu\)-integrable) and the right-hand side of \((61)\).
A.3.1 Smoothness computations

Here we provide detailed computations for the smoothness coefficients used in Lemma 9. We assume here that the assumptions A1, A2, and A3 are satisfied. Starting with the difference of expected gradients, using Jensen’s inequality and the smoothness assumption A2, we have

$$\|E_\mu [L'(h_1) - L'(h_2)]\| \leq E_\mu \|L'(h_1) - L'(h_2)\| \leq \lambda_1 \|h_1 - h_2\|. \quad (64)$$

As discussed in §4.3, differences of gradients modulated by $\rho'$ are slightly more complicated. In particular, recalling the equality (40), the norm of the difference

$$E_\mu \rho'_\sigma (L(h_1) - \theta_1)L'(h_1) - E_\mu \rho'_\sigma (L(h_2) - \theta_2)L'(h_2) \quad (65)$$

can be bounded above by the sum of

$$E_\mu \|L'(h_1)\|\|\rho'_\sigma (L(h_1) - \theta_1) - \rho'_\sigma (L(h_2) - \theta_2)\| \quad (66)$$

and

$$E_\mu |\rho'_\sigma (L(h_2) - \theta_2)||L'(h_1) - L'(h_2)||. \quad (67)$$

We take up (66) and (67) one at a time. Starting with (66), from A3 we know that the dispersion derivative $\rho'$ is $\lambda_2$-Lipschitz, and thus we have

$$(66) \leq \left( \frac{\lambda_2}{\sigma} \right) E_\mu \|L'(h_1)\| (\|L(h_1) - L(h_2)\| + |\theta_1 - \theta_2|)$$

$$\leq \left( \frac{\lambda_2}{\sigma} \right) E_\mu \|L'(h_1)\| (\|h_1 - h_2\| \sup_{0 < c < 1} \|L'((1 - c)h_1 - ch_2)\| + |\theta_1 - \theta_2|)$$

$$\leq \left( \frac{\lambda_2}{\sigma} \right) (\|h_1 - h_2\| E_\mu \|L'\|_H^2 + |\theta_1 - \theta_2| \sup_{h \in H} E_\mu \|L'(h)\|)$$

$$\leq \lambda_3 (\|h_1 - h_2\| + |\theta_1 - \theta_2|). \quad (68)$$

Here, the second inequality uses the helper Lemma 13 and our assumption of differentiability, while the third inequality uses our assumption on the expected squared norm of the gradient. We have set the Lipschitz coefficient $\lambda_3$ in (68) to be

$$\lambda_3 := \left( \frac{\lambda_2}{\sigma} \right) \left[ E_\mu \|L'\|_H^2 + \sup_{h \in H} E_\mu \|L'(h)\| \right].$$

This gives us a bound on (66). Moving on to (67), if $\rho'$ is bounded on $\mathbb{R}$, then we have

$$(67) \leq |\rho'|_\infty E_\mu \|L'(h_1) - L'(h_2)\|$$

$$\leq \lambda_4 \|h_1 - h_2\| \quad (69)$$

with $\lambda_4 := \lambda_1 |\rho'|_\infty$, recalling the bound (64). To summarize, we can use (68) and (69) to control (65) as follows:

$$(65) \leq (66) + (67)$$

$$\leq \lambda_3 (\|h_1 - h_2\| + |\theta_1 - \theta_2|) + \lambda_4 \|h_1 - h_2\|$$

$$\leq \lambda_5 (\|h_1 - h_2\| + |\theta_1 - \theta_2|)$$

where $\lambda_5 := \lambda_3 + \lambda_4$. With these preparatory details organized, it is straightforward to obtain a Lipschitz property on the gradients of $r_\mu(\cdot)$ and $\tilde{r}_\mu(\cdot)$, as summarized in Lemma 9, and detailed in the proof below.
Proof of Lemma 9. Take any two pairs \((h_1, \theta_1), (h_2, \theta_2) \in \mathcal{H} \times \mathbb{R}\), and for readability let us write \(r_{p,j} := r_p(h_j, \theta_j)\) and \(\tilde{r}_{p,j} := \tilde{r}_p(h_j, \theta_j)\) for \(j \in \{1, 2\}\). Using our upper bounds on (64) and (65), recalling the form of (36), we have

\[
\|\partial_h r_{p,1} - \partial_h r_{p,2}\| \\
\leq \|E_\mu [L'(h_1) - L'(h_2)]\| + \left(\frac{\eta}{\sigma}\right) \|E_\mu \left[\rho' \left(\frac{L(h_1) - \theta_1}{\sigma}\right) L'(h_1) - \rho' \left(\frac{L(h_2) - \theta_2}{\sigma}\right) L'(h_2)\right]\| \\
\leq \lambda_1 \|h_1 - h_2\| + \left(\frac{\eta \lambda_5}{\sigma}\right) (\|h_1 - h_2\| + |\theta_1 - \theta_2|) \\
\leq \left(\lambda_1 + \left(\frac{\eta \lambda_5}{\sigma}\right)\right) (\|h_1 - h_2\| + |\theta_1 - \theta_2|).
\]

This can be done almost identically for \(\tilde{r}_p\), recalling the form (37) we have

\[
\|\partial_h \tilde{r}_{p,1} - \partial_h \tilde{r}_{p,2}\| \leq \left(\frac{\lambda_5}{\sigma}\right) (\|h_1 - h_2\| + |\theta_1 - \theta_2|).
\]

Next, let us look at the partial derivative taken with respect to the threshold parameter, recalling the forms (34) and (35). To bound the absolute value of these differences, using the generalized mean value theorem (Lemma 13) again, we have for \(r_p\) that

\[
|\partial_\theta r_{p,1} - \partial_\theta r_{p,2}| \leq \left(\frac{\eta}{\sigma}\right) \|E_\mu \left[\rho' \left(\frac{L(h_1) - \theta_1}{\sigma}\right) - \rho' \left(\frac{L(h_2) - \theta_2}{\sigma}\right)\right]\| \\
\leq \left(\frac{\eta \lambda_2}{\sigma^2}\right) \|E_\mu|L(h_1) - L(h_2)| + |\theta_1 - \theta_2|\| \\
\leq \left(\frac{\eta \lambda_2}{\sigma^2}\right) \|E_\mu\| H_r(\|h_1 - h_2\| + |\theta_1 - \theta_2|).
\]

As for \(\tilde{r}_p\), just note that \(|\partial_\theta r_{p,1} - \partial_\theta r_{p,2}| / \eta = |\partial_\theta \tilde{r}_{p,1} - \partial_\theta \tilde{r}_{p,2}|\). Taking the preceding upper bounds together, for \(r_p\) we have

\[
\|r'_p(h_1, \theta_1) - r'_p(h_2, \theta_2)\| = \|\partial_h r_{p,1} - \partial_h r_{p,2}\| + |\partial_\theta r_{p,1} - \partial_\theta r_{p,2}| \\
\leq \left(\lambda_1 + \left(\frac{\eta \lambda_5}{\sigma}\right) + \frac{\eta \lambda_2}{\sigma^2}\right) \|E_\mu\| H_r(\|h_1 - h_2\| + |\theta_1 - \theta_2|),
\]

noting that the initial equality follows from the fact that we are using the sum of norms for our product space norm here (see also Remark 10). Identically, for \(\tilde{r}_p\) we have

\[
\|\tilde{r}'_p(h_1, \theta_1) - \tilde{r}'_p(h_2, \theta_2)\| \leq \left(\frac{\lambda_5}{\sigma} + \frac{\lambda_2}{\sigma^2}\right) \|E_\mu\| H_r(\|h_1 - h_2\| + |\theta_1 - \theta_2|).
\]

These bounds on the gradient differences are precisely the desired result.

\[
\square
\]

**B Supplementary technical facts**

**B.1 Lipschitz properties**

Here we give a fundamental property of differentiable functions that generalizes the mean value theorem.

**Lemma 13.** Let \(\mathcal{U}\) and \(\mathcal{V}\) be normed linear spaces, and let \(f: \mathcal{U} \to \mathcal{V}\) be Fréchet differentiable on an open set \(S \subseteq \mathcal{U}\). Taking any \(u \in S\), we have

\[
\|f(u + u') - f(u)\| \leq \|u'\| \sup_{0 < c < 1} \|f'(u + cu')\|
\]

for any \(u' \in \mathcal{U}\) such that \(u + cu' \in S\) for all \(0 \leq c \leq 1\).
Proof. See Luenberger [37, §7.3, Prop. 2].

Note that Lemma 13 has the following important corollary: *bounded gradients imply Lipschitz continuity*. In particular, if \( \| f'(u) \| \leq \lambda < \infty \) for all \( u \in S \), then it follows immediately that \( f \) is \( \lambda \)-Lipschitz on \( S \).

A closely related result goes in the other direction. Let \( f : U \to \mathbb{R} \) be convex and \( \lambda \)-Lipschitz. If \( f \) is sub-differentiable at a point \( u \in U \), then we have

\[
|\langle \partial f(u), u' - u \rangle| \leq |f(u') - f(u)| \leq \lambda \| u' - u \|.
\]

As such, for convex, sub-differentiable functions, \( \lambda \)-Lipschitz continuity implies that all sub-gradients are bounded as \( \| \partial f(x) \| \leq \lambda \).

### B.2 Convexity

**Lemma 14.** Let function \( f : \mathbb{R}^d \to \mathbb{R} \) be twice continuously differentiable. Then \( f \) is convex if and only if its Hessian is positive semi-definite, namely when

\[
(f''(v)u, u) \geq 0
\]

for all \( u, v \in \mathbb{R}^d \).

**Proof.** See Nesterov [43, Thm. 2.1.4].

### B.3 Derivatives for the Barron class

Let \( \rho(\cdot; \alpha) \) be defined according to (5). Here we compute derivatives of the map \( x \mapsto \rho_\sigma(x; \alpha) \), using the shorthand notation \( \rho_\sigma(x; \alpha) := \rho(x/\sigma; \alpha) \). We denote the first derivative of \( \rho_\sigma(\cdot; \alpha) \) evaluated at \( x \in \mathbb{R} \) by \( \rho'_\sigma(x; \alpha) \), which is computed as

\[
\rho'_\sigma(x; \alpha) = \begin{cases} 
\frac{x}{\sigma^2}, & \text{if } \alpha = 2 \\
\frac{2x}{x^2 + 2\sigma^2}, & \text{if } \alpha = 0 \\
\frac{(x/\sigma)^2 \exp\left(-\frac{(x/\sigma)^2}{2}\right) - \frac{1}{\sigma} (x/\sigma)^2}{\left(1 + \frac{(x/\sigma)^2}{|\alpha-2|}\right)^{\alpha/2}}, & \text{if } \alpha = -\infty \\
\frac{x}{\sigma^2} \left(1 + \frac{(x/\sigma)^2}{|\alpha-2|}\right)^{\alpha/2} - 1 \left(1 - \frac{1-|\alpha|}{|\alpha-2|} \frac{2(x/\sigma)^2}{1+(x/\sigma)^2/|\alpha-2|}\right), & \text{otherwise.}
\end{cases}
\] (70)

In the same way, letting \( \rho''_\sigma(x; \alpha) \) denote the second derivative of \( \rho_\sigma(\cdot; \alpha) \) evaluated at \( x \in \mathbb{R} \), this is computed as

\[
\rho''_\sigma(x; \alpha) = \begin{cases} 
\frac{1}{\sigma^2}, & \text{if } \alpha = 2 \\
\frac{2}{x^2 + 2\sigma^2} \left(1 - \frac{2\sigma^2}{x^2 + 2\sigma^2}\right), & \text{if } \alpha = 0 \\
\frac{1}{\sigma^2} \left(1 + \frac{(x/\sigma)^2}{|\alpha-2|}\right)^{\alpha/(\alpha-2)} - 1 \left(1 - \frac{1-|\alpha|}{|\alpha-2|} \frac{2(x/\sigma)^2}{1+(x/\sigma)^2/|\alpha-2|}\right), & \text{if } \alpha = -\infty \\
\frac{1}{\sigma^2} \left(1 + \frac{(x/\sigma)^2}{|\alpha-2|}\right)^{\alpha/(\alpha-2)} \left(1 - \frac{1-|\alpha|}{|\alpha-2|} \frac{2(x/\sigma)^2}{1+(x/\sigma)^2/|\alpha-2|}\right), & \text{otherwise.}
\end{cases}
\] (71)

We emphasize that \( \rho'_\sigma(x; \alpha) \) and \( \rho''_\sigma(x; \alpha) \) are not equal to \( \rho'(x/\sigma; \alpha) \) and \( \rho''(x/\sigma; \alpha) \), but by a simple application of the chain rule are easily seen to satisfy the relations

\[
\rho'_\sigma(x; \alpha) = \frac{1}{\sigma} \rho'(x/\sigma; \alpha), \quad \rho''_\sigma(x; \alpha) = \frac{1}{\sigma^2} \rho''(x/\sigma; \alpha)
\]

for any \( x \in \mathbb{R}, \sigma > 0, \) and \( \alpha \in [-\infty, 2] \).
B.4 Elementary inequalities

The following elementary inequalities will be of use.

\[
\left(1 + \frac{x}{p}\right)^p \geq \left(1 + \frac{x}{q}\right)^q, \quad \forall x \geq 0, \quad p > q > 0
\]  
(72)

\[
(1 + x)^c \leq 1 + \frac{cx}{1 - (c - 1)x}, \quad -1 \leq x < \frac{1}{c - 1}, \quad c > 1
\]
(73)

The inequality below is sometimes referred to as Bernoulli’s inequality.

\[
(1 + x)^a \geq 1 + ax, \quad \forall x > -1, \quad a \geq 1.
\]
(74)

B.5 Expected dispersion is coercive

**Lemma 15** (Expected dispersion is coercive). Let \( f : \mathbb{R} \to \mathbb{R}_+ \) be any non-negative function which is even (i.e., \( f(x) = f(-x) \) for all \( x \in \mathbb{R} \)) and non-decreasing on \( \mathbb{R}_+ \). Let \( X \) be any random variable such that \( E_\mu f(X - \theta) < \infty \) for all \( \theta \in \mathbb{R} \). Then, we have

\[
\lim_{|\theta| \to \infty} E_\mu f(X - \theta) = \lim_{x \to \infty} f(x)
\]
and note that this includes the divergent case where \( f(x) \to \infty \) as \( |x| \to \infty \).

**Proof of Lemma 15.** By our assumptions, we have \( f(x) \geq 0 \) and \( f(-x) = f(x) \) for all \( u \in \mathbb{R} \), and \( f(x_1) \leq f(x_2) \) whenever \( 0 \leq x_1 \leq x_2 \). With these facts in place, note that for any choice of \( a \geq 0 \) and \( \theta \) such that \( |\theta| \geq a \), we have

\[
E_\mu f(X - \theta) = E_\mu f(|\theta - X|)
\geq E_\mu f(|||\theta| - |X|||)
\geq E_\mu I_{\{|X| \leq a\}} f(|||\theta| - |X|||)
\geq f(||\theta| - a) P\{|X| \leq a\}.
\]
(75)

For readability, let us write \( f(\pm \infty) \) for the limit of \( f(x) \) as \( |x| \to \infty \). Trivially, we know that \( E_\mu f(X - \theta) \leq f(\pm \infty) \). Using the preceding inequality (75), we have a lower bound of \( E_\mu f(X - \theta) \geq f(\pm \infty) P\{|X| \leq a\} \) that holds for any \( a \geq 0 \). When \( f(\pm \infty) = \infty \), the desired result is immediate. When \( f(\pm \infty) < \infty \), simply note that \( \{\{|X| \leq a\} \uparrow \Omega \} \) as \( a \uparrow \infty \), and thus using the continuity of probability measures, we have \( P\{|X| \leq a\} \to 1 \) as \( a \to \infty \). Thus, the lower bound (75) can be taken arbitrarily close to \( f(\pm \infty) \), implying the desired result. \( \square \)

C Benchmark dataset information

We use eight well-known benchmark datasets in this paper, and in our figures we identify them respectively by the following keywords: adult,\(^{35}\) cifar10,\(^{36}\) cod_rna,\(^{37}\) covtype,\(^{38}\) emnist_balanced,\(^{39}\) fashion_mnist,\(^{40}\) mnist,\(^{41}\) and protein.\(^{42}\) Further background on all datasets is available at the URLs provided in the footnotes.

\(^{34}\) All countably additive set functions on \( \sigma \)-fields satisfy such continuity properties [3, Thm. 1.2.7].

\(^{35}\) https://archive.ics.uci.edu/ml/datasets/Adult

\(^{36}\) https://www.cs.toronto.edu/~kriz/cifar.html

\(^{37}\) https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

\(^{38}\) https://archive.ics.uci.edu/ml/datasets/covtype

\(^{39}\) https://www.nist.gov/itl/products-and-services/emnist-dataset

\(^{40}\) https://github.com/zalandoresearch/fashion-mnist

\(^{41}\) http://yann.lecun.com/exdb/mnist/

\(^{42}\) https://www.kdd.org/kdd-cup/view/kdd-cup-2004/Data