Machine Learning (2022) 111:4679–4718
https://doi.org/10.1007/s10994-022-06217-5

Learning with risks based on M-location

Matthew J. Holland[1]

Received: 6 December 2021 / Revised: 13 April 2022 / Accepted: 2 July 2022 /
Published online: 25 August 2022
© The Author(s) 2022

Abstract
In this work, we study a new class of risks defined in terms of the location and deviation of the loss distribution, generalizing far beyond classical mean-variance risk functions. The class is easily implemented as a wrapper around any smooth loss, it admits finite-sample stationarity guarantees for stochastic gradient methods, it is straightforward to interpret and adjust, with close links to M-estimators of the loss location, and has a salient effect on the test loss distribution, giving us control over symmetry and deviations that are not possible under naive ERM.

Keywords Risk-sensitive learning · Non-convex non-smooth risk minimization · Stochastic optimization

1 Introduction

In machine learning, the important yet ambiguous notion of "good off-sample generalization" (or "small test error") is typically formalized in terms of minimizing the expected value of a random loss $\mathbf{E}_{\mu}\,\mathrm{L}(h)$, where h is a candidate decision rule and $\mathrm{L}(h)$ is a random variable on an underlying probability space $(\Omega, \mathcal{F}, \mu)$. This setup based on *average* off-sample performance has been famously called the "general setting of the learning problem" by Vapnik (1999), and is central to the decision-theoretic formulation of learning in the influential work of Haussler (1992). This is by no means a purely theoretical concern; when average performance dictates the ultimate objective of learning, the data-driven *feedback* used for training in practice will naturally be designed to prioritize average performance (Bottou et al. 2016; Johnson and Zhang 2014; Le Roux et al. 2013; Shalev-Shwartz and Zhang 2013). Take the default optimizers in popular software libraries such as PyTorch or TensorFlow; virtually without exception, these methods amount to efficient implementations of empirical risk minimization. While the minimal expected loss formulation is clearly an intuitive choice, the tacit emphasis on average performance represents an

---

Editors: Krzysztof Dembczynski and Emilie Devijver.

✉ Matthew J. Holland
matthew-h@ar.sanken.osaka-u.ac.jp

[1] Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

important and non-trivial value judgment, which may or may not be appropriate for any given real-world learning task.

1.1 Our basic motivation

Put simply, our goal is to develop learning algorithms that make risk function design an explicit part of the modern machine learning workflow, rather than the traditional tacit emphasis of average performance while ignoring all other properties of the (test) loss distribution that are inherent in traditional empirical risk minimization (ERM). More concretely, we want to design a transformation that can be applied to *any* loss function, which is flexible enough to express varying degrees of sensitivity to loss deviations and tail behavior, easy to implement using existing stochastic gradient-based frameworks (e.g., PyTorch), and ultimately leads to novel risks which are analytically tractable. In the following paragraphs, we will discuss the basic ideas underlying our approach at a high level, and touch on some of the key related literature. After some basic theoretical setup, we discuss the strengths and drawbacks of our proposed approach (relative to the existing literature) in more detail in Sect. 2.3.

1.2 Some statistical context

In order to make the value judgement of "how to define risk" an explicit part of the machine learning methodology, in this work we consider a generalized class of risk functions, designed to give the user much greater flexibility in terms of how they choose to evaluate performance, while still allowing for theoretical performance guarantees. One core statistical concept is that of the *M-location* of the loss distribution under a candidate h, defined by

$$\mathrm{M}(h) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}} \; \mathbf{E}_{\mu} \rho\left(\frac{\mathrm{L}(h) - \theta}{\sigma}\right). \tag{1}$$

Here $\rho : \mathbb{R} \to [0, \infty)$ is a function controlling how we measure deviations, and $\sigma > 0$ is a scaling parameter. Since the loss distribution μ is unknown, clearly $\mathrm{M}(h)$ is an ideal, unobservable quantity. If we replace μ with the empirical distribution induced by a sample $(\mathrm{L}_1, \ldots, \mathrm{L}_n)$, then for certain special cases of ρ we get an *M-estimator* of the location of $\mathrm{L}(h)$, a classical notion dating back to Huber (1964), which justifies our naming. Ignoring integrability concerns for the moment, note that in the special case of $\rho(u) = u^2$, we get the classical risk $\mathrm{M}(h) = \mathbf{E}_{\mu} \mathrm{L}(h)$, and in the case of $\rho(u) = |u|$, we get $\mathrm{M}(h) = \inf\{u : \mu\{\mathrm{L}(h) \leq u\} \geq 0.5\}$, namely the median of the loss distribution. This rich spectrum of evaluation metrics makes the notion of casting learning problems in terms of minimization of M-locations (via their corresponding M-estimators) very conceptually appealing. However, while the statistical properties of the minima of M-estimators in special cases are understood (Brownlees et al. 2015), the optimization involved is both costly and difficult, making the task of designing and studying $\mathrm{M}(\cdot)$-minimizing learning *algorithms* highly intractable. With these issues in mind, we study a closely related alternative which retains the conceptual appeal of raw M-locations, but is more computationally congenial.

1.3 Our approach

With σ and ρ as before, our generalized risks will be defined implicitly by

$$\min_{\theta \in \mathbb{R}} \left[\theta + \eta \, \mathbf{E}_{\mu} \rho\left(\frac{\mathrm{L}(h) - \theta}{\sigma} \right) \right], \tag{2}$$

where $\eta > 0$ is a weighting parameter that controls the balance of priority between location and deviation. A more formal definition will be given in Sect. 2 (see Eqs. 3–5), including concrete forms for $\rho(\cdot)$ that are conducive to both fast computation and meaningful learning guarantees. In addition, we will show (cf. Proposition 3) that this new objective can be written as

$$[\mathrm{M}(h) - c_{\mathrm{M}}] + \eta \, \mathbf{E}_{\mu} \rho\left(\frac{\mathrm{L}(h) - [\mathrm{M}(h) - c_{\mathrm{M}}]}{\sigma} \right),$$

where the shift term $c_{\mathrm{M}} > 0$ can be simply characterized by the equality

$$\mathbf{E}_{\mu} \rho' \left(\frac{\mathrm{L}(h) - [\mathrm{M}(h) - c_{\mathrm{M}}]}{\sigma} \right) = \frac{\sigma}{\eta},$$

noting that $c_{\mathrm{M}} \to 0_{+}$ as $\eta \to \infty$. By utilizing smoothness properties of loss functions typical to machine learning problems (e.g., squared error, cross-entropy, etc.), even though the generalized risks need not be convex, they can be shown to satisfy weak notions of convexity, which still admit finite-sample guarantees of near-stationary for stochastic gradient-based learning algorithms (details in Sect. 3). This approach has the additional benefit that implementation only requires a single wrapper around any given loss which can be set prior to training, making for easy integration with frameworks such as PyTorch and TensorFlow, while incurring negligible computational overhead.

1.4 Our contributions

The key contribution here is a new concrete class of risk functions, defined and analyzed in Sect. 2. These risks are statistically easy to interpret, their empirical counterparts are simple to implement in practice, and as we prove in Sect. 3, their design allows for standard stochastic gradient-based algorithms to be given competitive finite-sample excess risk guarantees. We also verify empirically (Sect. 4) that the proposed feedback generation scheme has a demonstrable effect on the test distribution of the loss, which as a side-effect can be easily leveraged to outperform traditional ERM implementations, a result which is in line with early insights from Breiman (1999) and Reyzin and Schapire (2006) regarding the impact of the loss distribution on generalization. More broadly, the choice of *which risk to use* plays a central role in pursuit of increased transparency in machine learning, and our results represent a first step towards formalizing this process.

1.5 Relation to existing literature

With respect to alternative notions of "risk" in machine learning, perhaps the most salient example is conditional value-at-risk (CVaR) (Prashanth et al. 2020; Holland and Haress 2021a), namelythe expected loss conditioned on it exceeding a quantile at a pre-specified probability level. CVaR allows one to encode a sensitivity to extremely large losses, and admits convexity when the underlying loss is convex, though the conditioning often leaves the effective sample size very small. Other notions such as spectral risk and cumulative prospect theory (CPT) scores have also been considered (Bhat and Prashanth 2020; Holland and Haress 2021b; Khim et al. 2020; Leqi et al. 2019), but the technical difficulties involved with computation and analysis arguably outweigh the conceptual benefits of learning using such scores. These proposals can all be interpreted as "location" parameters of the underlying loss distribution; our risks take the form of a sum of a location and a deviation term, where the location is a shifted M-location, as described above. The basic notion of combining location and deviation information in evaluation is a familiar concept; the mean-variance objective $\mathbf{E}_{\mu} \mathrm{L}(\cdot) + \operatorname{var}_{\mu} \mathrm{L}(\cdot)$ dates back to classical work by Markowitz (1952); our proposed class includes this as a special case, but generalizes far beyond it. Mean-variance and other risk function classes are studied by Ruszczyński and Shapiro (2006, 2006b), who give minimizers a useful dual characterization, though our proposed class is not captured by their work (see also Remark 4). We note also that the recent (and independent) work of Lee et al. (2020) considers a form which is similar to (2) in the context of empirical risk minimizers; the critical technical difference is that their formulation is restricted to ρ which is monotonic, an assumption which enforces convexity. The special case of mean-variance is also treated in depth in more recent work by Duchi and Namkoong (2019), who consider stochastic learning algorithms for doing empirical risk minimization with variance-based regularization. Finally, we note that our technical analysis in Sect. 3 makes crucial use of weak convexity properties of function compositions, an area of active research in recent years (Duchi and Ruan 2018; Drusvyatskiy and Paquette 2019; Davis and Drusvyatskiy 2019). Since our proposed objective can be naturally cast as a composition taking us from parameter space to a Banach space and finally to $\mathbb{R}$, leveraging the insights of these previous works, we extend the existing machinery to handle learning over Banach spaces, and give finite-sample guarantees for arbitrary Hilbert spaces. More details are provided in Sect. 3, plus the appendix.

1.6 Notation and terminology

To give the reader an approximate idea of the technical level of this paper, we assume some familiarity with probability spaces, the notion of sub-gradients and the sub-differential of convex functions, as well as special classes of vector spaces like Banach and Hilbert spaces, although the main text is written with a wide audience in mind. Strictly speaking, we will also deal with sub-differentials of non-convex functions, but these technical concepts are relegated to the appendix, where all formal proofs are given. In the main text, to improve readability, we write $\partial f(x)$ to denote the sub-differential of f at x, regardless of the convexity of f. When we refer to a function being λ-*smooth*, this refers to its gradient being λ-Lipschitz continuous, and *weak* smoothness just requires such continuity on directional derivatives; all these concepts are given a detailed introduction in the appendix. Throughout this paper, we use $\mathbf{E}[\cdot]$ for taking expectation, and $\mathbf{P}$ as a general-purpose probability function. For indexing, we will write $[k] := \{1, \ldots, k\}$. Distance of a vector v from a set A will be denoted by $\operatorname{dist}(v;A) := \inf\{\|v - v'\| : v' \in A\}$.

2 A concrete class of risk functions

The risks described by (2) are fairly intuitive as-is, but a bit more structure is needed to ensure they are well-defined and useful in practice. In this section, we introduce a concrete choice of ρ for measuring deviations, consider a risk class modulated via re-scaling, establish basic theoretical properties and discuss both the utility and limitations of the proposed risk class.

2.1 Deviations and re-scaling

To make things more concrete, let us fix ρ as

$$\rho(u) := u\,\mathrm{atan}\,(u) - \frac{\log(1+u^2)}{2}, \quad u \in \mathbb{R}. \tag{3}$$

This function is handy in that it behaves approximately quadratically around zero, and it is both $\pi/2$-Lipschitz and strictly convex on the real line.[1] Fixing this particular choice of ρ and letting Z be a random variable (any $\mathcal{F}$-measurable function), we interpolate between mean- and median-centric quantities via the following class of functions, indexed by $\sigma \in [0,\infty]$:

$$\mathrm{r}_{\sigma}(Z,\theta) := \theta + \eta\,\mathbf{E}_{\mu}\rho_{\sigma}(Z-\theta), \text{ where } \rho_{\sigma}(u) := \begin{cases} |u|, & \text{if } \sigma = 0 \\ \rho(u/\sigma), & \text{if } 0 < \sigma < \infty \\ u^2, & \text{if } \sigma = \infty. \end{cases} \tag{4}$$

With this class of ancillary functions in hand, it is natural to define

$$\mathrm{R}_{\sigma}(Z) := \inf_{\theta\in\mathbb{R}} \mathrm{r}_{\sigma}(Z,\theta) \tag{5}$$

to construct a class of *risk functions*. In the context of learning, we will use this risk function to derive a generalized *risk*, namely the composite function $h \mapsto \mathrm{R}_{\sigma}(\mathrm{L}(h))$. As a special case, clearly this includes risks of the form (2) given earlier. Visualizations of ρ highlighting the role of re-scaling are given in Fig. 1. See also Fig. 5 in the appendix for visualizations of different loss functions transformed according to (4). Minimizing $\mathrm{R}_{\sigma}(\mathrm{L}(h))$ in h is our formal learning problem of interest.

2.2 Basic properties

Before considering learning algorithms, we briefly cover the basic properties of the functions r_{σ} and R_{σ}. Without restricting ourselves to the specialized context of "losses," note that if Z is any square-μ-integrable random variable, this immediately implies that $|\mathrm{r}_{\sigma}(Z,\theta)| < \infty$ for all $\theta \in \mathbb{R}$, and thus $\mathrm{R}_{\sigma}(Z) < \infty$. Furthermore, the following result shows that it is straightforward to set the weight η to ensure $\mathrm{R}_{\sigma}(Z) > -\infty$ also holds, and a minimum exists.

---

[1] To see this, note that $\rho'(u) = \mathrm{atan}\,(u)$ and $\rho'(\mathbb{R}) \subset (-\pi/2, \pi/2)$, and $\rho''(u) = 1/(1+u^2) > 0$.

Proposition 1 *Assuming that* $\mathbf{E}_{\mu}Z^{2} < \infty$, *set* η *based on* $\sigma \in [0,\infty]$ *as follows*: *if* $\sigma = 0$, *take* $\eta > 1$; *if* $0 < \sigma < \infty$, *take* $\eta > 2\sigma/\pi$; *if* $\sigma = \infty$, *take any* $\eta > 0$. *Under these settings, the function* $\theta \mapsto \mathrm{r}_{\sigma}(Z,\theta)$ *is bounded below and takes its minimum on* $\mathbb{R}$. *Thus, for each square-*μ*-integrable* Z, *there always exists a* (*non-random*) $\theta_{Z} \in \mathbb{R}$ *such that*

$$\mathrm{R}_{\sigma}(Z) = \theta_{Z} + \eta\,\mathbf{E}_{\mu}\rho_{\sigma}\big(Z - \theta_{Z}\big). \tag{6}$$

Furthermore, when $\sigma > 0$, *this minimum* θ_{Z} *is unique.*

Remark 2 *(Mean-median interpolation)* In order to ensure that risk modulation via $\sigma \in [0,\infty]$ smoothly transitions from a median-centric ($\sigma = 0$ case) to a mean-centric ($\sigma = \infty$ case) location, the parameter η plays a key role. Noting that for any $u \in \mathbb{R}$, for ρ defined by (3) we have $2\sigma^{2}\rho(u/\sigma) \to u^{2}$ as $\sigma \to \infty$, and thus for large values of $\sigma > 0$ it is natural to set $\eta = 2\sigma^{2}$. On the other end of the spectrum, since $\sigma\log(1 + (u/\sigma)^{2}) \to 0_{+}$ whenever $\sigma \to 0_{+}$, it is thus natural to set $\eta = \sigma/\operatorname{atan}(\infty) = 2\sigma/\pi$ when $\sigma > 0$ is small. Strictly speaking, in light of the conditions in Proposition 1, to ensure R_{σ} is finite one should take $\eta > 2\sigma/\pi$.

What can we say about our risk functions R_{σ} in terms of more traditional statistical risk properties? The form of R_{σ} given in (6) has a simple interpretation as a weighted sum of "location" and "deviation" terms. In the statistical risk literature, the seminal work of Artzner et al. (1999) gives an axiomatic characterization of location-based risk functions that can be considered *coherent*, while Rockafellar et al. (2006) characterize functions which capture the intuitive notion of "deviation," and establish a lucid relationship between coherent risks and their deviation class. The following result describes key properties of the proposed risk functions, in particular highlighting the fact that while our location terms are monotonic, our risk functions are non-traditional in that they are non-monotonic.

Proposition 3 (Non-monotonic risk functions) *Let* $\mathcal{Z}$ *be a Banach space of square-*μ*-integrable functions. For any* $\sigma \in [0,\infty]$, *let* $\eta > 0$ *be set as in Proposition* 1. *Then, the functions* $\mathrm{r}_{\sigma} : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ *and* $\mathrm{R}_{\sigma} : \mathcal{Z} \to \mathbb{R}$ *satisfy the following properties*:

- Both r_{σ} and R_{σ} are continuous, convex, and sub-differentiable.
- The location in (6) is monotonic (i.e., $Z_1 \leq Z_2$ implies $\theta_{Z_1} \leq \theta_{Z_2}$) and translation-equivariant (i.e., $\theta_{Z+a} = \theta_{Z} + a$ for any $a \in \mathbb{R}$), for any $0 < \sigma \leq \infty$.
- The deviation in (6) is non-negative and translation-invariant, namely for any $a \in \mathbb{R}$, we have $\mathbf{E}_{\mu}\rho_{\sigma}(Z + a - \theta_{Z+a}) = \mathbf{E}_{\mu}\rho_{\sigma}(Z - \theta_{Z})$, for any $0 < \sigma \leq \infty$.
- The risk R_{σ} is not monotonic, i.e., $\mu\{Z_1 \leq Z_2\} = 1$ need not imply $\mathrm{R}_{\sigma}(Z_1) \leq \mathrm{R}_{\sigma}(Z_2)$.

In particular, the risk $h \mapsto \mathrm{R}_{\sigma}(\mathrm{L}(h))$ need not be convex, even if $\mathrm{L}(\cdot)$ is.

Remark 4 Since our risk function R_{σ} is not monotonic, standard results in the literature on optimizing generalized risks do not apply here. We remark that our proposed risk class does not appear among the comprehensive list of examples given in the works of Ruszczyński and Shapiro (2006, 2006b), aside from the special case of $\sigma = \infty$ with $\eta = 1$. While the continuity and sub-differentiability of any risk function which is convex *and* monotonic is well-known for a large class of Banach spaces (Ruszczyński and Shapiro 2006, Sec. 3), in Proposition 3 we obtain such properties without monotonicity by using square-μ-integrability combined with properties of our function class ρ_{σ}.

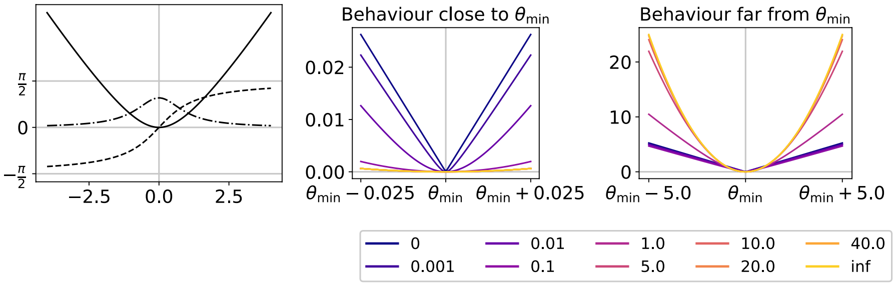


Fig. 1 Left: graphs of ρ defined in (3) (solid line), ρ' (dashed line), and ρ'' (dot-dash line). Middle–right: these are graphs of $\theta \mapsto \eta\rho_{\sigma}(1.0 - \theta)$, with ρ_{σ} defined in (4), where the minimizer is $\theta_{\min} = 1.0$, the colors correspond to different σ values, and η is set following Remark 2. That is, for $0 < \sigma < 1.0$, set $\eta = \sigma / \operatorname{atan}(\infty)$. For $\sigma = 0$, set $\eta = 1.05$. For $1.0 \leq \sigma < \infty$, set $\eta = 2\sigma^2$. For $\sigma = \infty$, set $\eta = 1.0$

Since our principal interest is the case where $Z = \mathrm{L}(h)$, the key takeaways from this section are that while the proposed risk $h \mapsto \mathrm{R}_{\sigma}(\mathrm{L}(h))$ is well-defined and easy to estimate given a random sample $\mathrm{L}_1(h), \ldots, \mathrm{L}_n(h)$, the learning task is non-trivial since $\mathrm{R}_{\sigma}(\mathrm{L}(\cdot))$ is not differentiable (and thus non-smooth) when $\sigma = 0$, and for any $\sigma \in [0, \infty]$ need not be convex, even when the underlying loss is both smooth and convex. Fortunately, smoothness properties of the losses typically used in machine learning can be leveraged to overcome these technical barriers, opening a path towards learning guarantees for practical algorithms, which is the topic of Sect. 3.

2.3 Strengths and limitations

Before diving further into the analysis of learning algorithms in the next section, let us pause for a moment to consider the following important question: *when should we use the proposed risk class over more traditional alternatives?* In addition, what kind of limitations or tradeoffs are faced when using this risk class? In the following paragraphs we attempt to provide an initial answer to these questions, in the context of traditional ERM and the alternative risk functions discussed in the literature review of Sect. 1.

2.3.1 Flexible control over deviations

The most obvious feature of the risk class defined in (4)–(5) is that it offers significant control over how we penalize *deviations* in the loss distribution. Unlike traditional ERM which simply requires that we minimize the location $\mathbf{E}_{\mu} \mathrm{L}(h)$, the proposed risk can only be made small when both the location and deviations are sufficiently small. It is well known that encouraging the (test) loss distribution to have small variance yields sharper bounds on the expected test loss than are available with naive ERM (Maurer and Pontil 2009). Furthermore, loss deviations are deeply linked to notions of *fairness* in machine learning (Williamson and Menon 2019), where an explicit design decision is made to ensure performance is similar across sensitive features (such as age, race, or gender), potentially at the cost of a larger expected loss. The alternative risks considered in this line of research are characterized by risks that add a deviation term to the expected loss $\mathbf{E}_{\mu} \mathrm{L}(h)$ (Hashimoto et al. 2018), and thus are always at least as sensitive to the loss distribution tails as the mean itself is. On the other hand, our deviation penalty based on σ and ρ can be used to enforce fairness with significant flexibility over the influence that errant observations have, since the location term is not fixed to the mean, but rather determined by the ρ and σ setting that we use; e.g., by taking $\sigma \to 0$ we obtain a location close to the median (often much lower than the mean), with deviations measured using a function that is insensitive to outliers. Please see Sect. 4 for empirical evidence that our risk class can be effectively used to control test loss deviations.

2.3.2 Symmetry

Another important feature of the risk class that we study here is the bidirectional nature of the function ρ used to measure deviations. This is in stark contrast with existing alternative risk classes such as CVaR, entropic risk, and other so-called "optimal certainty equivalent" risks (Lee et al. 2020), as well as distributionally robust optimization (DRO) risks (Namkoong and Duchi 2016; Zhai et al. 2021), which all place a strong emphasis on the loss tails on the upside, while downplaying or completely ignoring tails on the downside. This difference becomes important for losses that are unbounded below and can take on negative values, such as in logistic regression or for more general negative log-likelihood minimization. When the loss distribution (during training) is asymmetric with heavy tails on the downside, risks such as CVaR will provide a much smaller penalty than our risk class, which by design picks up on deviations in both directions. This can be expected to encourage greater symmetry in the test loss distribution, a phenomenon that we have almost observed empirically (e.g., Fig. 4).

2.3.3 Drawbacks and workarounds

One of the key limitations to our approach is that the risk class proposed does not preserve convexity when the underlying loss $\mathrm{L}(h)$ is convex in h. One obvious limitation is that analysis of learning algorithms cannot in general yield (global) excess risk bounds, but rather must be limited to either analysis centered around an arbitrary local minimum or global analysis of stationarity (see Sect. 3 for more details). Another side to this is computational. For extremely computationally intensive tasks where production-class convex solvers are an essential ingredient, our risk class cannot be used as-is. One natural work-around is to replace $\rho(\cdot)$ with $\rho((\cdot)_{+})$, which sacrifices the aforementioned symmetry of the deviation measure in exchange for monotonicity and convexity, while still retaining flexible control over the influence of tails on the upside. On the other hand, a great deal of modern machine learning applications are built upon non-linear models (e.g., most neural networks) which involve $\mathrm{L}(h)$ that is non-convex in h to begin with, so the drawbacks discussed here really only arise within the context of (large production-grade) linear models.

3 Learning algorithm analysis

Algorithm 1 Projected sub-gradient method with randomized output.

```
inputs: initial point $(h_0, \theta_0) \in C \subset \mathcal{H} \times \mathbb{R}$, step sizes $(\alpha_t)$, and max iterations $n$.

for $t \in \{0, \ldots, n-1\}$ do
    Sample $G_t$ via (9).
    Update $(h_t, \theta_t) \mapsto (h_{t+1}, \theta_{t+1})$ via (8).
end for

Sample $T \in \{0, \ldots, n-1\}$ with probabilities $\mathbf{P}\{T = t\} = \alpha_t / (\sum_{k=0}^{n-1} \alpha_k)$, $t \in [n-1]$.

return: $(\bar{h}_{[n]}, \bar{\theta}_{[n]}) := (h_T, \theta_T)$.
```

Thus far, we have only been concerned with *ideal* quantities R_{σ} and r_{σ} used to define the ultimate formal goal of learning. In practice, the learner will only have access to noisy, incomplete information. In this work, we focus on iterative algorithms based on stochastic gradients, largely motivated by their practical utility and ubiquity in modern machine learning applications. For the rest of the paper, we overload our risk definitions to enhance readability, writing $\mathrm{r}_{\sigma}(h, \theta) := \mathrm{r}_{\sigma}(\mathrm{L}(h), \theta)$ and $\mathrm{R}_{\sigma}(h) := \mathrm{R}_{\sigma}(\mathrm{L}(h))$. First note that we can break down the underlying joint objective as $\mathrm{r}_{\sigma}(h, \theta) = \mathbf{E}_{\mu}(f_2 \circ F_1)(h, \theta)$, where we have defined

$$F_1(h, \theta) := (\mathrm{L}(h), \theta), \qquad f_2(u, \theta) := \theta + \eta \rho_{\sigma}(u - \theta). \tag{7}$$

From the point of view of the probability space $(\Omega, \mathcal{F}, \mu)$, the function F_1 is random, whereas f_2 is deterministic; our use of upper- and lower-case letters is just meant to emphasize this. Given some initial value $(h_0, \theta_0) \in \mathcal{H} \times \mathbb{R}$, one naively hopes to construct an efficient stochastic gradient algorithm using the update

$$(h_{t+1}, \theta_{t+1}) = \Pi_C\big[(h_t, \theta_t) - \alpha_t G_t\big], \tag{8}$$

where $\alpha_t > 0$ is a step-size parameter, $\Pi_C[\cdot]$ denotes projection to some set $C \subset \mathcal{H} \times \mathbb{R}$, and the stochastic feedback G_t is just a composition of sub-gradients, namely

$$G_t \in \partial f_2(\mathrm{L}(h_t), \theta_t) \circ \partial F_1(h_t, \theta_t). \tag{9}$$

We call this approach "naive" since it is exactly what we would do if we knew *a priori* that the underlying objective was convex and/or smooth.[2] The precise learning algorithm studied here is summarized in Algorithm 1. Fortunately, as we describe below, this naive procedure actually enjoys lucid non-asymptotic guarantees, on par with the smooth case.

---

[2] The convex and smooth case is the classical setting (Nemirovsky and Yudin 1983); see Shalev-Shwartz and Ben-David (2014) for a modern textbook introduction. When convex but non-smooth, see Shamir and Zhang (2013). When smooth but non-convex, see Ghadimi and Lan (2013).

3.1 How to measure algorithm performance?

Before stating any formal results, we briefly discuss the means by which we evaluate learning algorithm performance. Since the sequence $(\mathrm{R}_{\sigma}(h_t))$ cannot be controlled in general, a more tractable problem is that of finding a *stationary point* of r_{σ}, namely any (h^*, θ^*) such that $0 \in \partial \mathrm{r}_{\sigma}(h^*, \theta^*)$. However, it is not practical to analyze $\operatorname{dist}(0;\partial \mathrm{r}_{\sigma}(h_t, \theta_t))$ directly, due to a lack of continuity. Instead, we consider a smoothed version of r_{σ}:

$$\widetilde{\mathrm{r}}_{\sigma,\beta}(h,\theta) := \inf_{h',\theta'} \left[\mathrm{r}_{\sigma}(h',\theta') + \frac{1}{2\beta}\|(h,\theta) - (h',\theta')\|^2 \right]. \tag{10}$$

This is none other than the *Moreau envelope* of r_{σ}, with weighting parameter $\beta > 0$. A familiar concept from convex analysis on Hilbert spaces (Bauschke and Combettes 2017, Ch. 12 and 24), the Moreau envelope of non-smooth functions satisfying weak convexity properties has recently been shown to be a very useful metric for evaluating stochastic optimizers (Davis and Drusvyatskiy 2019; Drusvyatskiy and Paquette 2019). Our basic performance guarantees will first be stated in terms of the gradient of the smoothed function $\widetilde{\mathrm{r}}_{\sigma,\beta}$. We will then relate this to the joint risk r_{σ} and subsequently the risk R_{σ}.

3.2 Guarantees based on joint risk minimization

Within the context of the stochastic updates characterized by (8)–(9), we consider the case in which $\mathcal{H}$ is any Hilbert space. All Hilbert spaces are reflexive Banach spaces, and the stochastic sub-gradient $G_t \in (\mathcal{H} \times \mathbb{R})^*$ (the dual of $\mathcal{H} \times \mathbb{R}$) can be uniquely identified with an element of $\mathcal{H} \times \mathbb{R}$, for which we use the same notation G_t. Denoting the partial sequence $G_{[t]} := (G_0, \ldots, G_t)$, we formalize our assumptions as follows:

A1. For all $h \in \mathcal{H}$, the random loss $\mathrm{L}(h)$ is square-μ-integrable, locally Lipschitz, and weakly λ-smooth, with a gradient satisfying $\mathbf{E}_{\mu}|\mathrm{L}'(h;\cdot)|^2 < \infty$.
A2. $\mathcal{H}$ is a Hilbert space, and $C \subset \mathcal{H} \times \mathbb{R}$ is a closed convex set.
A3. The feedback (9) satisfies $\mathbf{E}[G_t \mid G_{[t-1]}] = \mathbf{E}_{\mu} G_t$ for all $t > 0$.[3]
A4. For some $0 < \kappa < \infty$, the second moments are bounded as $\mathbf{E}_{\mu}\|G_t\|^2 \leq \kappa^2$ for all t.

The following is a performance guarantee for Algorithm 1 in terms of the smoothed joint risk.

Theorem 5 (Nearly-stationary point of smoothed objective) *If* $0 < \sigma < \infty$, *set smoothing parameter* $\gamma = (1 + \eta\pi/(2\sigma))\max\{1, \lambda\}$. *Otherwise, if* $\sigma = 0$, *set* $\gamma = (1 + \eta)\max\{1, \lambda\}$. *Under these σ-dependent settings and assumptions A1–A4, let* $(\overline{h}_{[n]}, \overline{\theta}_{[n]})$ *denote the output of Algorithm* 1, *with* $\mathrm{r}_C^* := \inf\{\mathrm{r}_{\sigma}(h,\theta) : (h,\theta) \in C\}$ *denoting the minimum over the feasible set and* $\Delta_0 := \widetilde{\mathrm{r}}_{\sigma,\beta}(h_0, \theta_0) - \mathrm{r}_C^*$ *denoting the initialization error. Then, for any choice of* $n > 1, \eta > 0$, *and* $\beta < 1/\gamma$, *we have that*

---

[3] The expectation on the left-hand side is with respect to the joint distribution of $G_{[t]}$ conditioned on $G_{[t-1]}$.

$$\mathbf{E}\,\|\widetilde{\mathrm{r}}'_{\sigma,\beta}(\overline{h}_{[n]},\overline{\theta}_{[n]})\|^2 \leq \left(\frac{1}{1-\beta\gamma}\right)\frac{\Delta_0+\gamma\kappa^2\sum_{t=0}^{n-1}\alpha_t^2/2}{\sum_{t=0}^{n-1}\alpha_t},$$

where expectation is taken over all the feedback $G_{[n-1]}$.

Remark 6 *(Sample complexity)* Let us briefly describe a direct take-away from Theorem 5. If Δ_0, γ, and κ are known (upper bounds will of course suffice), then constructing step sizes as $\alpha_t^2 \geq \Delta_0/(n\gamma\kappa^2)$, if we set $\beta = 1/(2\gamma)$, it follows immediately that

$$\mathbf{E}\,\|\widetilde{\mathrm{r}}'_{\sigma,\beta}(\overline{h}_{[n]},\overline{\theta}_{[n]})\|^2 \leq \sqrt{\frac{2\gamma\kappa^2\Delta_0}{n}}.$$

Fixing some desired precision level of $\sqrt{\mathbf{E}\,\|\widetilde{\mathrm{r}}'_{\sigma,\beta}(\overline{h}_{[n]},\overline{\theta}_{[n]})\|^2} \leq \varepsilon$, the sample complexity is $\mathcal{O}(\varepsilon^{-4})$. This matches guarantees available in the smooth (but non-convex) case (Ghadimi and Lan 2013), and suggests that the "naive" strategy implemented by Algorithm 1 in fact comes with a clear theoretical justification.

3.3 Implications in terms of the original objective

The results described in Theorem 5 and Remark 6 are with respect to a smoothed version of the joint risk function r_σ. Linking these facts to insights in terms of the original proposed risk R_σ can be done as follows. Assuming we take $n \geq 2\gamma\kappa^2\Delta_0/\varepsilon^4$ to achieve the ε -precision discussed in Remark 6, the immediate conclusion is that the algorithm output is $(\varepsilon/2\gamma)$-close to a ε-nearly stationary point of r_σ. More precisely, we have that there exists an ideal point (h_n^*, θ_n^*) such that

$$\mathbf{E}\left[\operatorname{dist}\left(0;\partial\,\mathrm{r}_\sigma(h_n^*,\theta_n^*)\right)\right] \leq \varepsilon, \text{ and } \mathbf{E}\left\|(\overline{h}_{[n]},\overline{\theta}_{[n]})-(h_n^*,\theta_n^*)\right\| = \frac{\varepsilon}{2\gamma}. \tag{11}$$

The above fact follows from basic properties of the Moreau envelope (cf. Appendix 2.4). These non-asymptotic guarantees of being close to a "good" point extend to the function values of the risk R_σ since we are close to a candidate h_n^* whose risk value can be no worse than

$$\mathbf{E}\left[\mathrm{R}_\sigma(h_n^*)\right] \leq \mathbf{E}\left[\mathrm{r}_\sigma(h_n^*,\theta_n^*)\right] \leq \mathbf{E}\left[\mathrm{r}_\sigma(\overline{h}_{[n]},\overline{\theta}_{[n]})\right].$$

We remark that these learning guarantees hold for a class of risks that are in general non-convex and need not even be differentiable, let alone satisfy smoothness requirements.

3.4 Key points in the proof of Theorem 5

Here we briefly highlight the key sub-results involved in proving Theorem 5; please see Appendix 3.2 for all the details. The key structure that we require is a smooth loss, reflected in assumption A1. This along with the Lipschitz property of our function ρ_σ for all $0 \leq \sigma < \infty$ allows us to prove that the underlying objective r_σ is *weakly convex*, where $\mathcal{H}$ can be any Banach space (Proposition 12); this generalizes a result of (Drusvyatskiy and Paquette 2019, Lem. 4.2) from Euclidean space to any Banach space. This alone is not enough to obtain the desired result. Note that the assumption A3 is very weak, and trivially satisfied in most traditional machine learning settings (e.g., where losses are based on a sequence of iid data points). The question of whether the feedback is *unbiased* or not, i.e., whether $\mathbf{E}_{\mu}G_t$ is in the sub-differential of r_{σ} at step t or not, is something that needs to be formally verified. In Proposition 14 we show that as long as the gradient has a finite expectation, this indeed holds for the feedback generated by (9), when $\mathcal{H}$ is any Banach space. With the two key properties of a weakly convex objective and unbiased random feedback in hand, we can leverage the techniques used in (Davis and Drusvyatskiy 2019, Thm. 3.1) for proximal stochastic gradient methods applied to weakly convex functions on $\mathbb{R}^d$, extending their core argument to the case of any Hilbert space. Combining this technique with the proof of weak convexity and unbiasedness lets us obtain Theorem 5.

4 Empirical analysis

In this section we introduce representative results for a series of experiments designed to investigate the quantitative and qualitative repercussions of modulating the underlying risk function class. We have prepared a GitHub repository that includes code for both re-creating the empirical tests and pre-processing all the benchmark data.[4]

4.1 Sanity check in one dimension

As a natural starting point, we use a toy example to ensure that Algorithm 1 takes us where we expect to go for a particular risk setting. Consider a loss on $\mathbb{R}$ with the form $\mathrm{L}(h) = h\,\mathrm{L}_{\mathrm{wide}} + (1-h)\,\mathrm{L}_{\mathrm{thin}}$, where $\mathrm{L}_{\mathrm{wide}}$ and $\mathrm{L}_{\mathrm{thin}}$ are random variables independent of h and each other. As a simple example, we use a folded Normal distribution for both, namely $\mathrm{L}_{*} = |\mathrm{Normal}(a_{*}, b_{*}^2)|$, where $a_{\mathrm{wide}} = 0$, $a_{\mathrm{thin}} = 2.0$, $b_{\mathrm{wide}} = 1.0$, and $b_{\mathrm{thin}} = 0.1$. For simplicity, we fix $\alpha_t = 0.001$ throughout, and each step uses a mini-batch of size 8.[5] Regarding the risk settings, we look in particular at the case of $\sigma = \infty$, where we modify the setting of $\eta = 2^k$ over $k = 0, 1, \ldots, 7$. Results averaged over 100 trials are given in Fig. 2. By modifying η, we can control whether the learning algorithm "prefers" candidates whose losses have a high degree of dispersion centered around a good location, or those whose losses are well-concentrated near a weaker location.

4.2 Impact of risk choice on linear regression

Next we consider how the key choice of σ (and thus the underlying risk R_{σ}) plays a role on the behavior of Algorithm 1. As another simple, yet more traditional example, consider linear regression in one dimension, where $Y = w_0^* + w_1^* X + \epsilon$, where X and ϵ are independent zero-mean random variables, and $(w_0^*, w_1^*) \in \mathbb{R}^2$ are unknown to the learner. Using squared error $\mathrm{L}(h) = (Y - h(X))^2$, we run Algorithm 1 again with mini-batches of size 8

---

[4] https://github.com/feedbackward/mrisk

[5] We remark that here and throughout our experiments we use this mini-batch size, but it is by no means critical to the analysis. It was set completely arbitrarily as an illustrative example, and the overall picture remains the same with other mini-batch sizes, up to proper step size settings, which can be easily tested using the aforementioned GitHub repository.

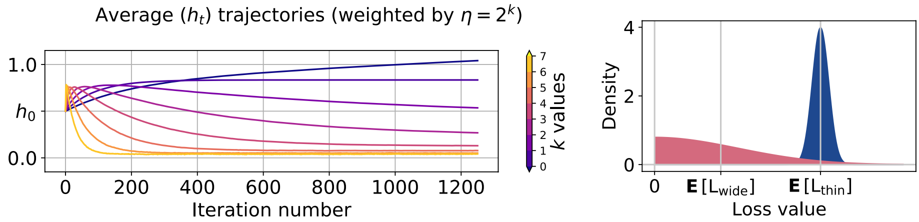


Fig. 2 A simple toy example using $\mathrm{L}(h) = h\,\mathrm{L}_{\text{wide}} + (1-h)\,\mathrm{L}_{\text{thin}}$. Trajectories shown are the sequence (h_t) generated by running (8) on $\mathbb{R}^2$, with $h_0 = 0.5$ and $\theta_0 = 0.5$, averaged over all trials. Densities of L_{wide} (red) and L_{thin} (blue) are also plotted, with additional details in the main text (Color figure online)

and $\alpha_t = 0.001$ fixed throughout, over a range of $\sigma \in [0, \infty]$ settings, for the same number of iterations as in the previous experiment. The initial value (h_0, θ_0) is initialized at zero plus uniform noise on $[-0.05, 0.05]$. We also consider multiple noise distributions; as a concrete example, letting $N = \text{Normal}(0, (0.8)^2)$, we consider both $\varepsilon = N$ (Normal case) and $\varepsilon = \mathrm{e}^N - \mathbf{E}\,\mathrm{e}^N$ (log-Normal case). In Fig. 3, we plot the learned regression lines (averaged over 100 trials) for each choice of σ and each noise setting. By modulating the target risk function, we can effectively choose between a self-imposed bias (smaller slope, lower intercept here), and a sensitivity to outlying values.

4.3 Tests using real-world data

Finally, we consider an application to some well-known benchmark datasets for classification. At a high level, we run Algorithm 1 for multi-class logistic regression for 10 independent trials, where in each trial we randomly shuffle and re-split each full dataset (88% training, 12% testing), and randomly re-initialize the model weights identically to the previous paragraph, again with mini-batches of size 8, and step sizes fixed to $\alpha_t = 0.01/\sqrt{d}$, where d is the number of free parameters. Additional background on the datasets is given in Appendix 6. The key question of interest is *how the test loss distribution changes* as we modify the learner's feedback to optimize a range of risks R_σ. In Fig. 4, we see a stark difference between doing traditional empirical risk minimization (ERM, denoted "off") and using R_σ-based feedback, particularly for moderately large values of σ. The logistic losses are concentrated much more tightly (visible in the bottom row histograms), and this also leads to a better classification error (visible in the top row plots), an interesting trend that we observed across many distinct datasets.

5 Concluding remarks

Moving forward, an appealing direction is to consider risk function classes which can encode properties going beyond location and deviation, such as explicit symmetry, multimodality, tail properties, and so forth. The hope is to develop a flexible and highly expressive framework coupled with efficient and practical algorithms with guarantees, which encodes distributional properties that are not readily captured by the current framework or traditional risk classes such as spectral risks (Khim et al. 2020).

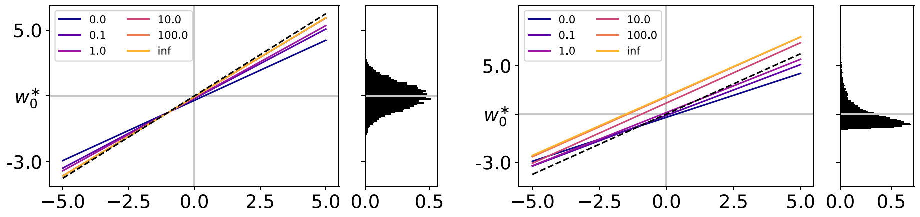


Fig. 3 Learned regression lines (solid; colors denote $\sigma \in [0, \infty]$) are plotted along with the true model $(w_0^*, w_1^*) = (1.0, 1.0)$ (dashed; black). Histograms are of independent samples of $w_0^* + \epsilon$. The left plots are the Normal case, and right plots are the log-Normal case

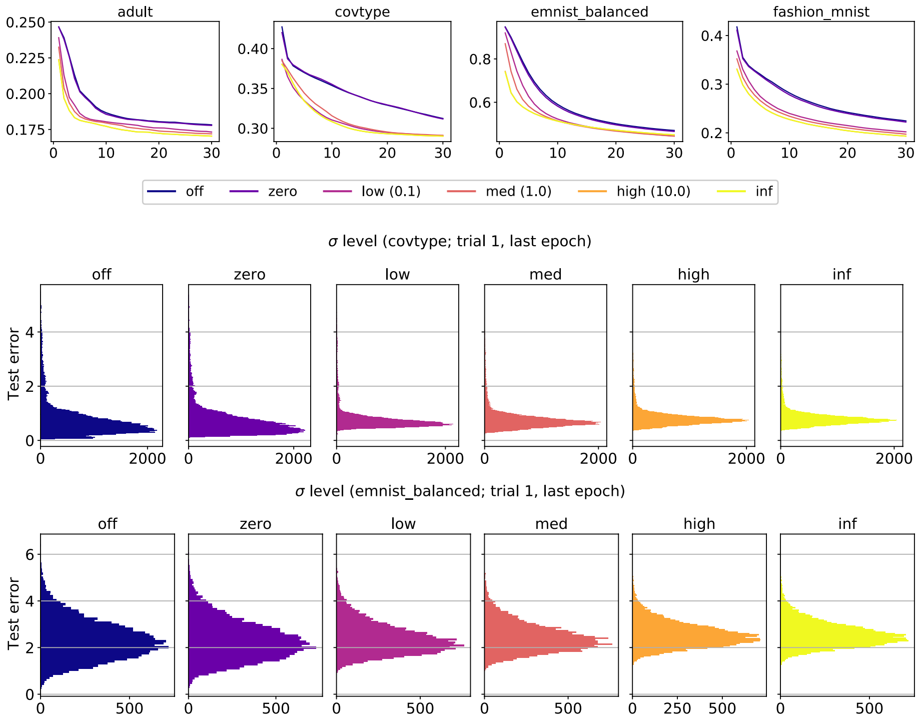


Fig. 4 Top row: average test error (zero-one loss) as a function of epochs, for four datasets and five σ levels, plus traditional ERM (denoted "off"). Middle and bottom rows: histograms of the test error (logistic loss) incurred after the final epoch for one trial under the `covtype` and `emnist_balanced` datasets

Taking the results of this paper together as a whole, we have obtained a generalized class of risk functions and a practical class of learning algorithms, which together effectively allow the user to encode the intuitive notions of *location* and *deviation* into the learning process. Of particular note is how changing the underlying risk has a lucid impact on the outcome of learning. In Sect. 4, we empirically verified that a simple modification of the scaling parameter of the underlying risk class (changing σ) was enough to result in a salient effect on the test loss distribution, using real-world datasets and typical stochastic gradient-based learning methods. While the importance of considering more properties of the loss distribution than just the

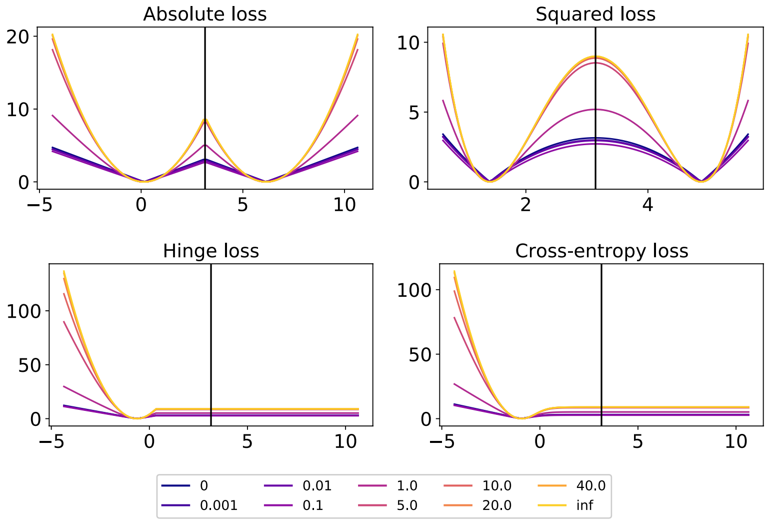


Fig. 5 Graphs of $h \mapsto \eta\rho_{\sigma}(\mathrm{L}(h) - \theta)$, with $h \in \mathbb{R}$, over different choices of σ and L, with $\theta = 3.0$ fixed, and η set as in Fig. 1. Colors correspond to σ, and each plot corresponds to a choice of $\mathrm{L}(\cdot)$. Absolute: $\mathrm{L}(h) = |h - h^{*}|$. Squared: $\mathrm{L}(h) = (h - h^{*})^{2}$. Hinge: $\mathrm{L}(h) = \max\{0, 1 - hh^{*}\}$. Cross-entropy: $\mathrm{L}(h) = \log(1 + \exp(-hh^{*}))$. In all cases, we have fixed $h^{*} = \pi$

expected value has been long understood (Breiman 1999; Reyzin and Schapire 2006), the key takeaway here is that we have seen how principled algorithmic modifications can bring about interpretable and meaningful effects on the test distribution, without having to abandon formal performance guarantees.

Appendix 1 Overview of appendix contents

Our appendix is comprised of several sections, ordered as follows:

2 Background and setup
3 Detailed proofs
4 Helper results
5 Loss example
6 Empirical supplement

As with the main paper, we handle theoretical topics before diving into empirical topics. Section 2 gives a very detailed background including numerous formal definitions, supporting lemmas, and discussion on results used later in the detailed proofs (Sect. 3) for the main paper's results. Additional numerical test results are at the very end of Sect. 6.

To provide additional visual intuition for the reader, we include several figures related to ρ defined in (3), ρ_{σ} defined in (4), and the resulting risk functions. In Fig. 1 we plot ρ and its derivatives, plus ρ_{σ} for a wide variety of $\sigma \in [0, \infty]$ values. Additional details are given in the figure caption. In Fig. 5, we show how specific choices of standard loss functions lead two different forms of the function composition $h \mapsto \mathrm{L}(h) \mapsto \eta\rho_{\sigma}(\mathrm{L}(h) - \theta)$.

Appendix 2 Background and setup

2.1 Preliminaries

General notation (probability) Underlying all our analysis is a probability space $(\Omega, \mathcal{F}, \mu)$.[6] All random variables, unless otherwise specified, will be assumed to be $\mathcal{F}$-measurable functions with domain Ω. Integration using μ will be denoted by $\mathbf{E}_{\mu} Z := \int_{\Omega} Z(\omega)\, \mu(\mathrm{d}\omega)$, and $\mathbf{P}$ will be used as a generic probability function, typically representing μ itself, or the product measure induced by a sample of random variables on $(\Omega, \mathcal{F}, \mu)$. We use $\mathcal{L}_2 := \mathcal{L}_2(\Omega, \mathcal{F}, \mu)$ to denote the set of all square-μ-integrable functions.[7]

General notation (normed spaces) Let $\mathcal{V}$ denote an arbitrary vector space. When we call $\mathcal{V}$ a normed (linear) space, we are referring to $(\mathcal{V}, \|\cdot\|)$, where $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ denotes the relevant norm. For any normed space $\mathcal{V}$, we shall denote by $\mathcal{V}^*$ the usual dual space of $\mathcal{V}$, namely all continuous linear functionals defined on $\mathcal{V}$. The space $\mathcal{V}^*$ is equipped with the norm $\|v^*\| := \sup\{v^*(u) : \forall\, u \in \mathcal{V}, \|u\| \leq 1\}$. We shall use the notation $\langle \cdot, \cdot \rangle$ to represent the "coupling" function between $\mathcal{V}$ and $\mathcal{V}^*$, that is for any $u \in \mathcal{V}$ and $v^* \in \mathcal{V}^*$, we will write $\langle u, v^* \rangle := v^*(u)$. For any sequence (x_n) of elements $x_1, x_2, \ldots \in \mathcal{V}$, we denote convergence of (x_n) to some element x' by $x_n \to x'$. When we take limits and do not specify a particular sequence, for example writing $x \to x'$, then this refers to any sequence (of elements from $\mathcal{V}$) that converges to x'. In the special case of real-valued sequences (where $\mathcal{V} \subset \mathbb{R}$), if we write $x_n \to x'_+$ (respectively $x_n \to x'_-$), this refers to all sequences from above (resp. below), i.e., any convergent sequence such that $x_n \geq x'$ (resp. $x_n \leq x'$) for all n. We denote the open ball of radius $r > 0$ centered at $x_0 \in \mathcal{V}$ by $B(x_0; r) := \{x \in \mathcal{V} : \|x_0 - x\| < r\}$. We denote the extended real line by $\overline{\mathbb{R}}$. On normed space $\mathcal{V}$, we denote the interior of a set $U \subset \mathcal{V}$ by $\operatorname{int} U$ (all $x \in U$ such that $B(x;\delta) \subset U$ for some δ).

General terminology On any normed linear space $\mathcal{V}$, a set $A \subset \mathcal{V}$ is said to be *compact* if for any sequence of elements in A, there exists a sub-sequence which converges on A. We denote the *effective domain* of an extended real-valued function f by $\operatorname{dom} f := \{x : f(x) < \infty\}$. We call a convex function $f : \mathcal{V} \to \overline{\mathbb{R}}$ *proper* if $f > -\infty$ and $\operatorname{dom} f \neq \emptyset$. We say that f is *coercive* if $\|x\| \to \infty$ implies $f(x) \to \infty$.[8] For a function $f : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{X}$ and $\mathcal{Y}$ being normed spaces, we say f is (locally) *Lipschitz* at $x_0 \in \mathcal{X}$ if there exists $\delta > 0$ and $\lambda > 0$ such that $x_1, x_2 \in B(x_0;\delta)$ implies $\|f(x_1) - f(x_2)\| \leq \lambda \|x_1 - x_2\|$. We say f is λ-Lipschitz on $\mathcal{X}$ if this property holds with a common coefficient λ for all $x_0 \in \mathcal{X}$.

Semi-continuous functions We say that a function f is *lower semi-continuous*[9] (LSC) at a point x if for each $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x - x'\| < \delta$ implies $f(x') > f(x) - \varepsilon$.

---

[6] For basic measure-theoretical facts supporting our main arguments, we use Ash and Doléans-Dade (2000) as a well-established and accessible reference. We will cite the exact results that pertain to our arguments in the main text as they become necessary.

[7] Strictly speaking, this is the set of all equivalence classes of square-μ-integrable functions, where $f \in \mathcal{L}_2(\Omega, \mathcal{F}, \mu)$ represents all functions that are equal μ-almost everywhere.

[8] For example, the function $f(x) = x^2$ is coercive, but $f(x) = \exp(x)$ is not.

[9] Nice references on semi-continuity: (Ash and Doléans-Dade 2000, Appendix 2), (Luenberger 1969, Ch. 2), (Barbu and Precupanu 2012, Sec. 2.1), (Penot 2012, Ch. 1).

If $-g$ is LSC, then we say g is *upper* semi-continuous (USC). The property that f is LSC at a point x is equivalent[10] to the property that for any sequence $x_n \to x$, we have

$$f(x) \leq \liminf_{n\to\infty} f(x_n). \tag{12}$$

Ordinary continuity is equivalent to being both USC and LSC, but the added generality of these weaker notions of continuity is often useful.

Differentiability We start by introducing some common notions of directional differentiability at a high level of generality.[11] Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces, $U \subset \mathcal{X}$ an open set, and $f : \mathcal{X} \to \mathcal{Y}$ a function of interest. The *radial derivative* of f at $x \in U$ in direction u is defined

$$f_r'(x;u) := \lim_{\alpha\to 0_+} \frac{f(x+\alpha u) - f(x)}{\alpha}. \tag{13}$$

A slight modification to this gives us the (Hadamard) *directional derivative* of f at $x \in U$ in direction u:

$$f'(x;u) := \lim_{(\alpha,u')\to(0_+,u)} \frac{f(x+\alpha u') - f(x)}{\alpha} \tag{14}$$

When $f_r'(x;u)$ exists for all directions u, we say that f is *radially differentiable* at x. Identically, when $f'(x;u)$ exists for all directions u, we say that f is *directionally differentiable* at x. When the map $u \mapsto f_r'(x;u)$ is continuous and linear, we say that f is *Gateaux differentiable* at x. When the map $u \mapsto f'(x;u)$ is continuous and linear, we say f is *Hadamard differentiable* at x. If f is Hadamard differentiable, then it is Gateaux differentiable. The converse does not hold in general, but if f is Lipschitz on a neighborhood of $x \in U$, then radial differentiability and directional differentiability (at x) are equivalent.[12]

When we simply say that a function $f : \mathcal{X} \to \mathcal{Y}$ is *differentiable* at $x \in U$, we mean that there exists a function $f'(x)(\cdot) : \mathcal{X} \to \mathcal{Y}$ that is linear, continuous, and which satisfies

$$\lim_{\|u\|\to 0} \frac{\|f(x+u) - f(u) - f'(x)(u)\|}{\|u\|} = 0. \tag{15}$$

This property is often referred to as *Fréchet differentiability*. When f is differentiable at x, the map $f'(x)$ is uniquely determined.[13] In the special case where $\mathcal{Y} \subset \mathbb{R}$, the linear functional represented by $f'(x) \in \mathcal{X}^*$ is called the *gradient* of f at x. Differentiability is also closely related to directional differentiability; if f is Gateaux differentiable on U and the map $x \mapsto f'(x;\cdot)$ is continuous at x, then f is differentiable at x.[14]

Sub-differentials Let $\mathcal{V}$ be any normed linear space. If $f : \mathcal{V} \to \overline{\mathbb{R}}$ is a (proper) convex function, the *sub-differential* of f at $x \in \operatorname{dom} f$ is defined

$$\begin{aligned}\partial f(x) &:= \left\{v^* \in \mathcal{V}^* : f(u) - f(x) \geq \langle u - x, v^*\rangle, u \in \mathcal{V}\right\}\\ &= \left\{v^* \in \mathcal{V}^* : f_r'(x;u) \geq \langle u, v^*\rangle, u \in \mathcal{V}\right\}.\end{aligned} \tag{16}$$

---

[10] (Ash and Doléans-Dade 2000, Thm. A2.2), (Penot 2012, Lem. 1.18).

[11] We follow basic notation and terminology of the authoritative text by Penot (2012).

[12] (Penot 2012, Prop. 2.25).

[13] See for example (Luenberger 1969, Ch. 7) or (Penot 2012, Ch. 2).

[14] (Penot 2012, Prop. 2.51).

The second characterization of $\partial f(x)$, given using the radial derivative (13), is useful and intuitive.[15] Some authors refer to this as the *Moreau-Rockafellar sub-differential* to emphasize the context of convex analysis.

More generally, however, if f is not convex, then the strong global property used to define the MR sub-differential is so restrictive that most interesting functions are left out. A more general notion is that of the *Fréchet sub-differential*.[16] Denoted $\partial_F f(x)$, the Fréchet sub-differential of f at x is the set of all bounded linear functionals $v^* \in \mathcal{V}^*$ such that for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|x-u\| \leq \delta \implies f(u)-f(x) \geq \langle u-x, v^* \rangle - \varepsilon \|x-u\|. \tag{17}$$

This local requirement is much weaker than the condition characterizing the MR-sub-differential, and clearly we have $\partial f(x) \subset \partial_F f(x)$. When f is assumed to be locally Lipschitz, another class of sub-differentials is often useful. Define the *Clarke directional derivative* of f at x in the direction u by

$$f_C'(x;u) := \limsup_{(\alpha,x')\to(0_+,x)} \frac{f(x'+\alpha u)-f(x')}{\alpha}. \tag{18}$$

The corresponding *Clarke sub-differential* is defined as

$$\partial_C f(x) := \{v^* \in \mathcal{V}^* : f_C'(x;u) \geq \langle u, v^* \rangle, u \in \mathcal{V}\}. \tag{19}$$

In the special case where f is convex, all the sub-differentials coincide, i.e., $\partial f(x) = \partial_F f(x) = \partial_C f(x)$.[17] We say that a function f is *sub-differentiable* at x if its sub-differential (in any sense) at x is non-empty. Finally, a remark on notation when using set-valued functions like $x \mapsto \partial_C f(x)$. When we write something like "we have $\langle u, \partial_C f(x) \rangle \geq g(u)$," it is the same as writing "we have $\langle u, v^* \rangle \geq g(u)$ for all $v^* \in \partial_C f(x)$." This kind of notation will be used frequently.

2.2 Generalized convexity

Let $\mathcal{X}$ be a normed linear space. Take an open set $U \subset \mathcal{X}$ and fix some point $x_0 \in U$. For a function $f : \mathcal{X} \to \overline{\mathbb{R}}$ and parameter $\gamma \in \mathbb{R}$, say that there exists $\delta > 0$ such that for all $x, x' \in B(x_0;\delta)$ and $\alpha \in (0,1)$, we have

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha) f(x') + \frac{\gamma\alpha(1-\alpha)}{2}\|x-x'\|^2. \tag{20}$$

When $\gamma \geq 0$, we say f is γ *-weakly convex* at x_0. When $\gamma \leq 0$, we say f is $(-\gamma)$ *-strongly convex* at x_0. When (20) holds for all $x_0 \in U$, we say that f is γ-weakly/strongly convex on U. The special case of $\gamma = 0$ is the traditional definition of convexity on U.[18]

The ability to construct a quadratic lower-bounding function for f is closely related to notions of weak/strong convexity. Consider the following condition: given $\gamma \in \mathbb{R}$, there exists $\delta > 0$ such that for all $x, x' \in B(x_0;\delta)$ we have

---

[15] See (Penot 2012, Thm. 3.22) for this fact.

[16] We follow (Penot 2012, Sec. 4.1) for terms and notation here.

[17] This follows from (Penot 2012, Prop. 5.34). See also (Penot 2012, Sec. 4.1.1, Exercise 1).

[18] See for example (Nesterov 2004, Ch. 3).

$$f(x') \geq f(x) + \langle x' - x, \partial_C f(x) \rangle - \frac{\gamma}{2} \|x - x'\|^2. \tag{21}$$

Here $\partial_C f$ denotes the Clarke sub-differential of f, defined by (19). Let us assume henceforth that $\mathcal{X}$ is Banach, f is locally Lipschitz, and $\partial_C f(x)$ is non-empty for all $x \in U$.[19] For any $\gamma \in \mathbb{R}$, it is straightforward to show that (20) $\Longrightarrow$ (21) holds.[20] Since (21) gives us a lower bound on both $f(x)$ and $f(x')$ for any x and x' close enough to x_0, adding up the inequalities immediately implies

$$\langle x - x', \partial_C f(x) - \partial_C f(x') \rangle \geq -\gamma \|x - x'\|^2. \tag{22}$$

When $\mathcal{X}$ is Banach and f is locally Lipschitz, it is straightforward to show that (22) $\Longrightarrow$ (20) is valid.[21] As such, for Banach spaces and locally Lipschitz functions, we have that the conditions (20), (21), and (22) are all equivalent for the general case of $\gamma \in \mathbb{R}$.

Let us consider one more closely related property on the same open set $U \subset \mathcal{X}$:

$$x \mapsto f(x) + \frac{\gamma}{2} \|x\|^2 \text{ is convex on } U. \tag{23}$$

In the special case where $\mathcal{X}$ is a real Hilbert space and the norm $\|\cdot\|$ is induced by the inner product as $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$, then for any $x, x' \in U$ and $\alpha \in \mathbb{R}$, the equality

$$\alpha \|x\|^2 + (1 - \alpha) \|x'\|^2 = \|\alpha x + (1 - \alpha) x'\|^2 + \alpha (1 - \alpha) \|x - x'\|^2 \tag{24}$$

is easily checked to be valid.[22] In this case, the equivalence (20) $\Longleftrightarrow$ (23) follows from direct verification using (24).[23]

The facts above are summarized in the following result:

Proposition 7 (Characterization of generalized convexity) *Consider a function* $f : \mathcal{X} \to \overline{\mathbb{R}}$ *on normed linear space* $\mathcal{X}$. *When* $\mathcal{X}$ *is Banach and* f *is locally Lipschitz, then with respect to open set* $U \subset \mathcal{X}$ *we have the following equivalence*:

(20) $\Leftrightarrow$ (21) $\Leftrightarrow$ (22).

When $\mathcal{X}$ *is Hilbert, this equivalence extends to* (23).

2.3 Function composition on normed spaces

Next we consider the properties of compositions involving functions which are smooth and/or convex. Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces. Let $g : \mathcal{X} \to \mathcal{Y}$ and $h : \mathcal{Y} \to \overline{\mathbb{R}}$ be the maps used in our composition, and denote by $f := h \circ g$ the composition, i.e., $f(x) = h(g(x))$ for each $x \in \mathcal{X}$. Our goal will be to present sufficient conditions for the composition f to be weakly convex on an open set $U \subset \mathcal{X}$, in the sense of (20). If we assume simply that h is convex, fixing any point $x_0 \in U$ such that h is sub-differentiable at $g(x_0)$, it follows that for any choice of $x \in \mathcal{X}$ we have

---

[19] (Penot 2012, Prop. 5.3).

[20] See for example (Daniilidis and Malick 2005, Thm. 3.1); in particular their proof of (i) $\Longrightarrow$ (iii). Their result is stated for $\mathcal{X} = \mathbb{R}^d$ and locally Lipschitz f, but the proof easily generalizes to Banach spaces. See also the remarks following their proof about how the local Lipschitz condition can be removed.

[21] Just apply the argument for (ii) $\Longrightarrow$ (i) employed by (Daniilidis and Malick 2005, Thm. 3.1), and strengthen their argument by using a more general form of Lebourg's mean value theorem (Penot 2012, Thm. 5.12).

[22] (Bauschke and Combettes 2017, Cor. 2.15).

[23] See also (Davis and Drusvyatskiy 2019, Lem. 2.1) for a similar result when $\mathcal{X} = \mathbb{R}^d$ and f is LSC.

$$f(x) = h(g(x)) \geq h(g(x_0)) + \langle g(x) - g(x_0), \partial h(g(x_0)) \rangle. \tag{25}$$

Let us further assume that h is λ_0-Lipschitz, and g is *smooth* in the sense that it is differentiable on U and the map $x \mapsto g'(x)$ is λ_1-Lipschitz. For readability, denote the derivative $g'(x_0) : \mathcal{X} \to \mathcal{Y}$ by $g'_0(\cdot) := g'(x_0)(\cdot)$. Taking any choice of $v_0 \in \partial h(g(x_0))$, we can write

$$\begin{aligned}
\langle g(x) - g(x_0), v_0 \rangle &= \langle g'_0(x - x_0), v_0 \rangle + \langle g(x) - g(x_0) - g'_0(x - x_0), v_0 \rangle \\
&\geq \langle g'_0(x - x_0), v_0 \rangle - \| g(x) - g(x_0) - g'_0(x - x_0) \| \| v_0 \| \\
&\geq \langle g'_0(x - x_0), v_0 \rangle - \frac{\lambda_1}{2} \| x - x_0 \|^2 \| v_0 \| \\
&\geq \langle g'_0(x - x_0), v_0 \rangle - \frac{\lambda_0 \lambda_1}{2} \| x - x_0 \|^2
\end{aligned} \tag{26}$$

The first inequality follows from the definition of the norm for linear functionals and the fact that $\partial h(g(x_0)) \subset \mathcal{Y}^*$. The second inequality follows from a Taylor approximation for Banach spaces (Proposition 16), using the smoothness of g. The final equality follows from the fact that for convex functions, the Lipschitz coefficient implies a bound on all sub-gradients, see (47). To deal with the remaining term, note that we can write

$$\langle g'_0(x - x_0), \partial h(g(x_0)) \rangle = \langle x - x_0, (g'_0)^*(\partial h(g(x_0))) \rangle = \langle x - x_0, \partial h(g(x_0)) \circ g'_0 \rangle. \tag{27}$$

To explain the notation here, we use $(\cdot)^*$ to denote the adjoint, namely $A^*(y^*) := y^* \circ A$, induced by any continuous linear map $A : \mathcal{X} \to \mathcal{Y}$, defined for each $y^* \in \mathcal{Y}^*$.[24] The special case we have considered here is where $Au = g'_0(u)$, noting that differentiability means that the map $u \mapsto g'_0(u)$ is continuous and linear. Recalling the desired form of (21), we need to establish a connection with $\partial_C f(x_0)$. If we further assume that g is locally Lipschitz, then we have

$$\partial_C f(x_0) \subset \partial_C h(g(x_0)) \circ g'_0 = \partial h(g(x_0)) \circ g'_0, \tag{28}$$

where the equality follows from the coincidence of sub-differentials in the convex case, and the key inclusion follows from direct application of a generalized chain rule.[25] Taking these facts together yields the following result.

Proposition 8 (Weak convexity for composite functions) *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. Let $g : \mathcal{X} \to \mathcal{Y}$ be locally Lipschitz and λ_1-smooth on an open set $U \subset \mathcal{X}$. Let $h : \mathcal{Y} \to \overline{\mathbb{R}}$ be convex and λ_0-Lipschitz on $g(U) \subset \mathcal{Y}$. Furthermore, let $g(U) \subset$ dom h. Then, the composite function $f := h \circ g$ is γ-weakly convex on U, for $\gamma = \lambda_0 \lambda_1$.*

---

[24] Written explicitly, for any $u \in \mathcal{X}$, we have $\langle u, A^*(y^*) \rangle = A^*(y^*)(u) = y^*(Au) = \langle Au, y^* \rangle$. If $A : \mathcal{X} \to \mathcal{Y}$ is linear and continuous, this implies that the adjoint A^* is a continuous linear map from $\mathcal{Y}^*$ to $\mathcal{X}^*$. For more general background: (Luenberger 1969, Ch. 6), (Penot 2012, Ch. 1).

[25] See for example (Penot 2012, Thm. 5.13(b)). This inclusion holds as long as g is strictly differentiable (Penot 2012, Defn. 2.54), a property implied by the smoothness we have assumed.

Proof The desired result just requires us to piece together the key facts we have outlined in the main text. Local Lipschitz properties for g and h imply that f is locally Lipschitz, and thus $\partial_C f(x_0) \neq \emptyset$ for all $x_0 \in U$. Using (28), we have that $\partial h(g(x_0)) \neq \emptyset$ as well. With this in mind, linking up (25)–(27), under the assumptions stated, for each $x, x_0 \in U$ we have

$$f(x) \geq f(x_0) + \langle x - x_0, \partial h(g(x_0)) \circ g_0' \rangle - \frac{\lambda_0 \lambda_1}{2} \|x - x_0\|^2. \tag{29}$$

Using the inclusion (28) with (29), we have (21) for $\gamma = \lambda_0 \lambda_1$, and the desired result holds since (21) implies (20). □

Remark 9 We note that Proposition 8 extends a result of (Drusvyatskiy and Paquette 2019, Lem. 4.2) from the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional Euclidean spaces, to the general Banach space setting considered here. For the classical case of Euclidean spaces, exact chain rules are well-known (Rockafellar and Wets 1998, Ch. 10.B).

2.4 Proximal maps of weakly convex functions

For normed linear space $\mathcal{X}$ and function $f : \mathcal{X} \to \overline{\mathbb{R}}$, the *Moreau envelope* $\operatorname{env}_\beta f$ and *proximal mapping* (or *proximity operator*) $\operatorname{prox}_\beta f$ are respectively defined for each $x \in \mathcal{X}$ as follows:

$$\operatorname{env}_\beta f(x) := \inf \left\{ f(x') + \frac{1}{2\beta} \|x - x'\|^2 : x' \in \mathcal{X} \right\} \tag{30}$$

$$\operatorname{prox}_\beta f(x) := \operatorname*{arg\,min}_{x' \in \mathcal{X}} \left[f(x') + \frac{1}{2\beta} \|x - x'\|^2 \right]. \tag{31}$$

Here $\beta > 0$ is a parameter. In the case where f is convex, the basic properties of the proximal map and envelope are well-understood, particularly when $\mathcal{X}$ is a Hilbert space.[26] These insights extend readily to the setting of weak convexity. Under the assumption that $\mathcal{X}$ is Hilbert, let f be γ-weakly convex on $\mathcal{X}$. Trivially we can write

$$f(x') + \frac{1}{2\beta} \|x - x'\|^2 = \left(f(x') + \frac{\gamma}{2} \|x - x'\|^2 \right) + \frac{1}{2} \left(\frac{1}{\beta} - \gamma \right) \|x - x'\|^2.$$

If we write $f_{\gamma,x}(u) := f(u) + (\gamma/2)\|x - u\|^2$ and $\beta_\gamma := (\beta^{-1} - \gamma)^{-1}$ for readability, then as long as $\beta_\gamma > 0$ we have for all $x \in \mathcal{X}$ that $\operatorname{env}_\beta f(x) = \operatorname{env}_{\beta_\gamma} f_{\gamma,x}(x)$ and $\operatorname{prox}_\beta f(x) = \operatorname{prox}_{\beta_\gamma} f_{\gamma,x}(x)$. By leveraging Proposition 7 under the Hilbert space assumption, we have that for any $x \in \mathcal{X}$, the function $f_{\gamma,x}(\cdot)$ is convex. This means that as long as $\beta_\gamma > 0$, namely whenever $\gamma < \beta^{-1}$, all the standard results available for the case of convex functions can be brought to bear on the problem.[27] Of particular importance to us is the fact that when f is LSC and γ-weakly convex, the Moreau envelope is differentiable, with gradient

---

[26] See for example (Bauschke and Combettes 2017, Ch. 12 and 24). For Banach spaces, modified notions of "proximity" measured using Bregman divergences have also been developed (Bauschke et al. 2003; Soueycatt et al. 2020). See also Jourani et al. (2014) for more analysis of the Moreau envelope in more general spaces.

[27] For example, see (Bauschke and Combettes 2017, Sec. 12.4).

$$(\operatorname{env}_{\beta} f)'(x) = \frac{1}{\beta}\left(x - \operatorname{prox}_{\beta} f(x)\right), \tag{32}$$

well-defined for all $\beta < \gamma^{-1}$ and $x \in \mathcal{X}$.[28] We will be interested in finding *stationary points* of f, namely those $x \in \mathcal{X}$ such that $0 \in \partial_C f(x)$. From the basic properties of the envelope and proximal mapping, for γ-weakly convex f we have

$$\operatorname{dist}(0; \partial_C f(\operatorname{prox}_{\beta} f(x))) \leq \|(\operatorname{env}_{\beta} f)'(x)\|. \tag{33}$$

That is, for any point $x \in \mathcal{X}$, the point $\operatorname{prox}_{\beta} f(x) \in \mathcal{X}$ is approximately stationary. The degree of precision is controlled by the gradient of $\operatorname{env}_{\beta} f$ evaluated at x. In addition, it follows immediately from (32) that

$$\|x - \operatorname{prox}_{\beta} f(x)\| = \beta \|(\operatorname{env}_{\beta} f)'(x)\|. \tag{34}$$

Since one trivially also has $f(\operatorname{prox}_{\beta} f(x)) \leq f(x)$, the norm of the gradient of $\operatorname{env}_{\beta} f$ evaluated at x also tells us how far we are from a point (namely $\operatorname{prox}_{\beta} f(x) \in \mathcal{X}$) which is no worse than x in terms of function value. These basic facts directly motivate the use of the Moreau envelope norm to quantify algorithm performance.[29]

Appendix 3 Detailed proofs

3.1 Proofs for Sect. 2

Lemma 10 (Lower semi-continuity) *Let $\mathcal{Z}$ be a linear space of $\mathcal{F}$-measurable random variables, and let $\rho : \mathbb{R} \to \overline{\mathbb{R}}$ be any non-negative LSC function that is Borel-measurable. Then we have that the functional $(Z, \theta) \mapsto \mathbf{E}_{\mu}\rho(Z - \theta)$ is also LSC.*

Proof of Lemma 10 Let (Z_k) and (θ_k) respectively be convergent sequences on $\mathcal{Z}$ and $\mathbb{R}$. As we take $k \to \infty$, say $Z_k \to Z_*$ pointwise, for some $Z \in \mathcal{Z}$, and $\theta_k \to \theta_* \in \mathbb{R}$. Since by assumption ρ is LSC on $\mathbb{R}$, using (12) we have (again, pointwise) that

$$\rho\left(Z_* - \theta_*\right) \leq \liminf_{k\to\infty} \rho\left(Z_k - \theta_k\right).$$

Writing $\rho_k := \rho(Z_k - \theta_k)$ for each $k \geq 1$ and $\rho_* = \rho(Z_* - \theta_*)$, it follows that

$$\mathbf{E}_{\mu}\rho_* \leq \mathbf{E}_{\mu}\left(\liminf_{k\to\infty} \rho_k\right) \leq \liminf_{k\to\infty} \left(\mathbf{E}_{\mu}\rho_k\right). \tag{35}$$

---

[28] One can just apply standard arguments such as given by (Bauschke and Combettes 2017, Prop. 12.30), while utilizing the weak convexity property described. See also (Drusvyatskiy and Paquette 2019, Lem. 4.3), (Davis and Drusvyatskiy 2019, Lem. 2.2), and (Poliquin and Rockafellar 1996, Thm. 4.4).

[29] This is highlighted in works such as Drusvyatskiy and Paquette (2019) and Davis and Drusvyatskiy (2019).

The former inequality follows from monotonicity of the integral, and the latter inequality follows from an application of Fatou's inequality, which is valid since $\rho_k \geq 0$.[30] Taking both ends of (35) together, since the choice of sequences (Z_k) and (θ_k) were arbitrary, it follows again from the equivalence (12) that the functional $(Z, \theta) \mapsto \mathbf{E}_{\mu}\rho((Z-\theta)/\sigma)$ is LSC on $\mathcal{Z} \times \mathbb{R}$. □

Lemma 11 (Basic integration properties) *Let* $\mathbf{E}_{\mu}Z^2 < \infty$ *hold, and take any* $\theta \in \mathbb{R}$. *Then the following properties of integrals based on* ρ_σ *defined in* (5) *hold*:

- For all $\sigma \in [0, \infty]$, we have $0 \leq \mathbf{E}_{\mu}\rho_\sigma(Z-\theta) < \infty$.
- For $0 < \sigma \leq \infty$, $\rho_\sigma(\cdot)$ is differentiable, and we have $\mathbf{E}_{\mu}|\rho_\sigma'(Z-\theta)| < \infty$.
- For $0 < \sigma \leq \infty$, $\rho_\sigma'(\cdot)$ is differentiable, and we have $0 \leq \mathbf{E}_{\mu}\rho_\sigma''(Z-\theta) < \infty$.

Furthermore, the Leibniz integration property holds for both derivatives, that is

$$\frac{d}{d\theta}\mathbf{E}_{\mu}\rho_\sigma(Z-\theta) = -\frac{\mathbf{E}_{\mu}\rho_\sigma'(Z-\theta)}{\sigma}, \qquad \frac{d}{d\theta}\mathbf{E}_{\mu}\rho_\sigma'(Z-\theta) = -\frac{\mathbf{E}_{\mu}\rho_\sigma''(Z-\theta)}{\sigma} \tag{36}$$

for any $0 < \sigma < \infty$, *and for the special case of* $\sigma = \infty$, *we have*

$$\frac{d}{d\theta}\mathbf{E}_{\mu}\rho_\infty(Z-\theta) = -2(\mathbf{E}_{\mu}Z-\theta), \qquad \frac{d}{d\theta}\mathbf{E}_{\mu}\rho_\infty'(Z-\theta) = -2. \tag{37}$$

These equalities hold for any $\theta \in \mathbb{R}$.

Proof of Lemma 11 Non-negativity of ρ_σ implies $0 \leq \mathbf{E}_{\mu}\rho_\sigma(Z-\theta)$ for all σ. Regarding finiteness, starting with $\sigma = 0$ we have $\mathbf{E}_{\mu}\rho_0(Z-\theta) = \mathbf{E}_{\mu}|Z-\theta| < \infty$, which follows from Hölder's inequality and μ-integrability of Z^2. For $0 < \sigma < \infty$, first note that $\operatorname{atan}'(u) \leq 1$ for all $u \in \mathbb{R}$, and thus since $\operatorname{atan}(0) = 0$, we have $|\operatorname{atan}(u)| \leq |u|$ and $u \operatorname{atan}(u) \leq u^2$ for all u. In particular, this means $\rho_\sigma(Z-\theta) \leq (Z-\theta)^2/\sigma^2$, and thus square-$\mu$-integrability of Z implies $\mathbf{E}_{\mu}\rho_\sigma(Z-\theta) < \infty$. The $\sigma = \infty$ case follows identically.

Moving to $\rho_\sigma'(\cdot)$, for $0 < \sigma < \infty$ we have that $|\rho_\sigma'(u)| < \pi/2$ for all $u \in \mathbb{R}$, and thus $\mathbf{E}_{\mu}|\rho_\sigma'(Z-\theta)| < \infty$. The exact same argument holds for $\rho_\sigma''(\cdot)$, since $0 < \rho_\sigma''(u) \leq 1$ for all $u \in \mathbb{R}$. The $\sigma = \infty$ case follows analogously.

For the Leibniz property, let (a_k) be any real sequence such that $a_k \to 0$. Using the fact that ρ_σ' is bounded, the dominated convergence theorem lets us deduce the following:

$$\begin{aligned}
\lim_{k\to\infty}\left[\frac{\mathbf{E}_{\mu}\rho_\sigma(Z-(\theta+a_k)) - \mathbf{E}_{\mu}\rho_\sigma(Z-\theta)}{a_k}\right] &= \lim_{k\to\infty}\mathbf{E}_{\mu}\left[\frac{\rho_\sigma(Z-(\theta+a_k)) - \rho_\sigma(Z-\theta)}{a_k}\right]\\
&= \mathbf{E}_{\mu}\left[\lim_{k\to\infty}\frac{\rho_\sigma(Z-(\theta+a_k)) - \rho_\sigma(Z-\theta)}{a_k}\right]\\
&= -\frac{\mathbf{E}_{\mu}\rho_\sigma'(Z-\theta)}{\sigma}.
\end{aligned}$$

We note that the first equality just uses μ-integrability and linearity of the Lebesgue integral, the second equality uses boundedness and integrability of the derivative, plus dominated convergence (e.g., (Ash and Doléans-Dade 2000, Thm. 1.6.9)). The last equality is just the chain rule applied to the differentiable function $\rho_\sigma(\cdot)$. Since the sequence (a_k) was arbitrary, we conclude that the first equality of (36) holds. The second equality of (36), as well as both equalities in (37) hold via an identical argument. □

---

[30] (Ash and Doléans-Dade 2000, Lem. 1.6.8).

Proof of Proposition 1 First, note that the convexity of $\rho_\sigma(\cdot)$ implies that $\theta \mapsto \mathrm{r}_\sigma(Z,\theta)$ is convex. We start by showing that r_σ is also coercive, namely that $|\theta| \to \infty$ implies $\mathrm{r}_\sigma(Z,\theta) \to \infty$. For all cases $\sigma \in [0,\infty]$, the non-negativity of ρ_σ and η trivially implies that $\theta \to \infty$ implies $\mathrm{r}_\sigma(Z,\theta) \to \infty$, and thus we need only consider the negative direction, where $\theta \to -\infty$.

For the case of $\sigma = 0$, note that

$$\theta + \eta\,\mathbf{E}_\mu|Z-\theta| \geq \theta - \eta|\theta| + \eta\,\mathbf{E}_\mu|Z|.$$

Clearly, with $\eta > 1$ the right-hand side grows without bound as $\theta \to -\infty$. For the case of $0 < \sigma = \infty$, writing $Z_\theta := (Z-\theta)/\sigma$ for readability, the joint risk can be written conveniently as

$$\mathrm{r}_\sigma(Z,\theta) = \theta\Big(1 - \frac{\eta}{\sigma}\,\mathbf{E}_\mu \operatorname{atan}\left(Z_\theta\right)\Big) + \frac{\eta}{\sigma}\,\mathbf{E}_\mu Z \operatorname{atan}\left(Z_\theta\right) - \frac{\eta}{2}\,\mathbf{E}_\mu \log\left(1+Z_\theta^2\right). \tag{38}$$

Since $\operatorname{atan}(\cdot)$ is monotonic (increasing) on $\mathbb{R}$, bounded as $|\operatorname{atan}(\cdot)| < \pi/2$, and $\operatorname{atan}(u) \to \pi/2$ as $u \to \infty$, we have that $\mathbf{E}_\mu \operatorname{atan}(Z_\theta) \to \pi/2$ as $\theta \to -\infty$, by monotone convergence.[31] Thus, taking $\eta > 2\sigma/\pi$ ensures that eventually as $\theta \to \infty$, the first term on the right-hand side of (38) will become positive. Since this term grows linearly, it dominates the other unbounded term (which is logarithmic), and thus we have shown that r_σ is coercive whenever $\sigma \geq 0$. Convexity and coercivity together imply that $\theta \mapsto \mathrm{r}_\sigma(Z,\theta)$ takes its minimum on $\mathbb{R}$; see (Bertsekas 2015, Sec. B.3.2) or (Barbu and Precupanu 2012, Thm. 2.11) for standard references.

The case of $\sigma = \infty$ is easy, since by direct inspection we can write

$$\mathrm{r}_\infty(Z,\theta) = \theta(1 - 2\eta\,\mathbf{E}_\mu Z) + \eta\theta^2 + \eta\,\mathbf{E}_\mu Z^2.$$

Since the sum of a strongly convex function and an affine function is strongly convex, we have that $\theta \mapsto \mathrm{r}_\infty(Z,\theta)$ has a unique minimum on $\mathbb{R}$.

It only remains to prove the uniqueness of θ_Z in the proposition statement for the case of $0 < \sigma < \infty$. The most direct way of doing this is to use the Leibniz property (36) proved in our helper Lemma 11, which in particular tells us that

$$\frac{d^2}{d\theta^2}\,\mathrm{r}_\sigma(Z,\theta) = \frac{\eta}{\sigma}\,\mathbf{E}_\mu \rho''\left(\frac{Z-\theta}{\sigma}\right) > 0,$$

where positivity follows from the fact that $\rho''(\cdot) = 1/(1+(\cdot)^2) > 0$. This implies strict convexity, and thus that the minimizer θ_Z is unique. □

Proof of Proposition 3 We take the points in the statement of the proposition in order, one at a time. To being, the (joint) convexity of r_σ follows from direct inspection, using the convexity of ρ_σ for any $\sigma \in [0,\infty]$. With this fact in mind, note that the convexity of R_σ can be checked easily as follows. For any $Z_1, Z_2 \in \mathcal{Z}$ and $\theta_1, \theta_2 \in \mathbb{R}$, the definition and convexity of r_σ immediately implies that for any $\alpha \in (0,1)$ we have

---

[31] (Ash and Doléans-Dade 2000, Thm. 1.6.7).

$$\begin{aligned}\mathrm{R}_\sigma(\alpha Z_1 + (1-\alpha)Z_2) &\leq \mathrm{r}_\sigma(\alpha Z_1 + (1-\alpha)Z_2, \alpha\theta_1 + (1-\alpha)\theta_2)\\ &\leq \alpha\,\mathrm{r}_\sigma(Z_1,\theta_1) + (1-\alpha)\,\mathrm{r}_\sigma(Z_2,\theta_2).\end{aligned}$$

Using the notation (and statement) of Proposition 1, we can set $\theta_1 = \theta_{Z_1} \in \mathbb{R}$ and $\theta_2 = \theta_{Z_2} \in \mathbb{R}$, and plugging this in to the above inequalities, we obtain

$$\mathrm{R}_\sigma(\alpha Z_1 + (1-\alpha)Z_2) \leq \alpha\,\mathrm{R}_\sigma(Z_1) + (1-\alpha)\,\mathrm{R}_\sigma(Z_2),$$

and thus both r_σ and R_σ are convex for any $\sigma \in [0,\infty]$. Note that this does not require the minima θ_{Z_1} and θ_{Z_2} to be unique, and thus holds for the $\sigma = 0$ case without issue. From Lemma 11, we also have that $|\,\mathrm{r}_\sigma(\cdot,\cdot)| < \infty$ and $|\,\mathrm{R}_\sigma(\cdot)| < \infty$, so both functions are proper convex. As for continuity, note that from Lemma 10 and the continuity of $\rho_\sigma(\cdot)$ for all $\sigma \in [0,\infty]$, we can immediately infer that r_σ is LSC. It is well-known that on Banach spaces, any proper convex LSC function is continuous and sub-differentiable on the interior of the effective domain.[32] Since our integrability assumptions imply $\operatorname{dom} \mathrm{r}_\sigma = \mathcal{Z} \times \mathbb{R}$, the continuity and sub-differentiability of r_σ is thus proved. To handle R_σ, take any sequence (Z_k) converging to an arbitrarily chosen point $Z_* \in \mathcal{Z}$. Let (θ_k) be any sequence converging to $\theta_{Z_*} \in \mathbb{R}$. Then by definition of R_σ and continuity of r_σ, we have

$$\limsup_{k\to\infty} \mathrm{R}_\sigma(Z_k) \leq \limsup_{k\to\infty} \mathrm{r}_\sigma(Z_k,\theta_k) = \lim_{k\to\infty} \mathrm{r}_\sigma(Z_k,\theta_k) = \mathrm{R}_\infty(Z_*,\theta_{Z_*}) = \mathrm{R}_\sigma(Z_*). \tag{39}$$

The two ends of the inequality (39) imply that R_σ is USC, via (12) and the relation of USC to LSC functions. On the effective domain of any convex USC function, the function is in fact continuous.[33] Thus, we have that R_σ is continuous. Furthermore, the sub-differentiability of R_σ follows in the exact same fashion as for r_σ.

Next, for the monotonicity of the location term $Z \mapsto \theta_Z$ in (6), with $0 < \sigma \leq \infty$, recall that we can utilize the Leibniz properties (36)–(37) from the integration Lemma 11. To start, we know that for any Z, the corresponding θ_Z must satisfy the following first-order optimality condition:

$$\text{If } \sigma < \infty: \quad \mathbf{E}_\mu \rho'_\sigma(Z - \theta_Z) = \frac{\sigma}{\eta}. \qquad \text{If } \sigma = \infty: \quad \theta_Z = \mathbf{E}_\mu Z - \frac{1}{2\eta}. \tag{40}$$

The desired monotonicity property is obvious for the $\sigma = \infty$ case using (40). As for the case of $0 < \sigma < \infty$, it is evident from the second equality of (36) that the function $\theta \mapsto \mathbf{E}_\mu \rho_\sigma(Z - \theta)$ is monotonically decreasing on $\mathbb{R}$. Thus, if we assume $Z_1 \leq Z_2$ almost surely but $\theta_{Z_1} > \theta_{Z_2}$, the first order optimality combined with monotonicity implies

$$\frac{\sigma}{\eta} = \mathbf{E}_\mu \rho_\sigma(Z_1 - \theta_{Z_1}) < \mathbf{E}_\mu \rho_\sigma(Z_1 - \theta_{Z_2}) \leq \mathbf{E}_\mu \rho_\sigma(Z_2 - \theta_{Z_2}) = \frac{\sigma}{\eta},$$

which is a contradiction. Thus, $\theta_{Z_1} \leq \theta_{Z_2}$ as desired for the $0 < \sigma < \infty$ case as well.

---

[32] Actually, via (Barbu and Precupanu 2012, Prop. 2.16), this holds for every point in the algebraic interior of its effective domain; the fact stated follows as the algebraic interior contains the interior.

[33] (Penot 2012, Prop. 3.2).

The translation-equivariance property of $Z \mapsto \theta_Z$ follows from direct inspection using the condition (40), that is for $0 < \sigma < \infty$, we trivially have

$$\frac{\sigma}{\eta} = \mathbf{E}_{\mu}\rho_{\sigma}(Z - \theta_Z) = \mathbf{E}_{\mu}\rho_{\sigma}(Z + a - (\theta_Z + a)),$$

and thus $\mathrm{r}_{\sigma}(Z + a, \theta_Z + a) = \mathrm{R}_{\sigma}(Z + a)$. Since Proposition 1 guarantees that the minimizer of $\theta \mapsto \mathrm{r}_{\sigma}(Z, \theta)$ is unique, we can safely write $\theta_{Z+a} = \theta_Z + a$. The proof for the $\sigma = \infty$ case is analogous.

Finally, to prove that R_{σ} is not in general monotonic, we give a concrete example of Z_1 and Z_2 such that $Z_1 \leq Z_2$ almost surely, but $\mathrm{R}_{\sigma}(Z_1) > \mathrm{R}_{\sigma}(Z_2)$. For simplicity, consider the case of $\sigma = \infty$, where $\rho_{\infty}(\cdot) = (\cdot)^2$, and for any $\eta > 0$ direct inspection shows that

$$\mathrm{R}_{\infty}(Z) = \mathbf{E}_{\mu}Z + \eta\,\mathrm{var}_{\mu}Z - \frac{1}{4\eta}. \tag{41}$$

That is, the special case of $\sigma = \infty$ is equivalent to the mean-variance risk function of classical portfolio theory, dating back to Markowitz (1952). The random variables Z_1 and Z_2 are constructed as follows. Let c_1 and c_2 be the respective centers, w_1 and w_2 the respective widths, and $v_1 < w_1^2$ and $v_2 < w_2^2$ the respective scaling factors of Z_1 and Z_2, which are characterized as

$$\mathbf{P}\left\{Z_j = c_j - w_j\right\} = \mathbf{P}\left\{Z_j = c_j + w_j\right\} = \frac{v_j}{2w_j^2}, \qquad \mathbf{P}\{Z_j = c_j\} = 1 - \mathbf{P}\{Z_j \neq c_j\}$$

for each $j \in \{1, 2\}$. Note that our assumptions imply $0 < \mathbf{P}\{Z_j = c_j\} < 1$, and direct inspection shows that $\mathbf{E}\,Z_j = c_j$ and $\mathrm{var}\,Z_j = v_j$, again for each $j \in \{1, 2\}$. As a simple concrete example, note that setting $c_2 = c_1 + w_1 + w_2$ guarantees $Z_1 \leq Z_2$ with probability 1. From the equality (41) given above, the difference in risks can be written as

$$\begin{aligned}\mathrm{R}_{\infty}(Z_1) - \mathrm{R}_{\infty}(Z_2) &= \mathbf{E}\,Z_1 - \mathbf{E}\,Z_2 + \eta\big(\mathrm{var}\,Z_1 - \mathrm{var}\,Z_2\big)\\ &= -(w_1 + w_2) + \eta(v_1 - v_2).\end{aligned}$$

Thus, $\mathrm{R}_{\infty}(Z_1) > \mathrm{R}_{\infty}(Z_2)$ holds whenever $v_1 - v_2 > w_1 + w_2$. For concreteness, say for some $\varepsilon > 0$, we fix the variance factors to $v_1 = w_1^2 - \varepsilon$ and $v_2 = w_2^2 - \varepsilon$ respectively. Then the condition simplifies to $w_1^2 > w_1 + w_2 + w_2^2$. As an example, setting $w_1 = 2$ and $w_2 = 1/2$, the condition holds, implying $\mathrm{R}_{\infty}(Z_1) > \mathrm{R}_{\infty}(Z_2)$, despite the fact that $Z_1 \leq Z_2$. This gives us a simple but intuitive example where monotonicity of R_{σ} does not hold, and concludes the proof. ☐

3.2 Proofs for Sect. 3

Recall that our basic probabilistic setup for the learning problem has an underlying probability space $(\Omega, \mathcal{F}, \mu)$, a hypothesis class $\mathcal{H}$, and a random loss $\mathrm{L}(h)$ indexed by $\mathcal{H}$. That is, we consider any $\mathcal{F}$-measurable function $\mathrm{L}(h;\cdot) : \Omega \to \mathbb{R}$ as a loss. When a particular realization $\omega \in \Omega$ is important, we will write $\mathrm{L}(h;\omega)$, but otherwise, for readability we will typically write $\mathrm{L}(h) := \mathrm{L}(h;\cdot)$. Our basic integrability assumption, carried over from Sect. 2, is that of square-μ-integrability, which in the context of losses is written explicitly as

$$\mathbf{E}_{\mu}|\mathrm{L}(h)|^2 = \int_{\Omega} |\mathrm{L}(h;\omega)|^2 \, \mu(\mathrm{d}\omega) < \infty$$

for all $h \in \mathcal{H}$. This requirement is made in assumption A1 in the main text. It follows that $\{\mathrm{L}(h) : h \in \mathcal{H}\} \subset \mathcal{L}_2(\Omega, \mathcal{F}, \mu)$. Thus the map $h \mapsto \mathrm{L}(h)$ takes us from $\mathcal{H}$ to $\mathcal{L}_2(\Omega, \mathcal{F}, \mu)$.

Loss-specific terminology To ensure our use of formal terms is clear, we apply the definitions of Sect. 2.1 to losses here. We shall typically suppress the dependence on $\omega \in \Omega$ in directional derivatives and gradients, writing $\mathrm{L}_r'(h;g) := \mathrm{L}_r'(h;g,\cdot)$, $\mathrm{L}'(h;g) := \mathrm{L}'(h;g,\cdot)$, and $\mathrm{L}'(h) := \mathrm{L}'(h;\cdot)$. Let $H \subset \mathcal{H}$ be an open set. We say that L *is radially differentiable* at $h \in H$ if the radial derivative $\mathrm{L}_r'(h;g)$ exists for all directions $g \in \mathcal{H}$, μ-almost surely. We say that L *is directionally differentiable* at $h \in H$ if the directional derivative $\mathrm{L}'(h;g)$ exists for all directions $g \in \mathcal{H}$, μ-almost surely. On this "good" event of probability 1, if the map $g \mapsto \mathrm{L}_r'(h;g)$ is linear and continuous, we say L *is Gateaux differentiable* at h, and if the map $g \mapsto \mathrm{L}'(h;g)$ is linear and continuous, we say L *is Hadamard differentiable* at h. When we say that L *is (Fréchet) differentiable* at $h \in H$, we mean that there exists a function $\mathrm{L}'(h)(\cdot) : \mathcal{H} \to \mathbb{R}_+$ that is linear, continuous, and which satisfies (15) μ-almost surely.[34] We say that L *is λ-Lipschitz at $h \in \mathcal{H}$* if there exists a $\delta > 0$ such that $\|h - h'\| < \delta \implies \|\mathrm{L}(h) - \mathrm{L}(h')\| \leq \lambda \|h - h'\|$. With the running assumption about second moments, this amounts to requiring

$$\|h - h'\| < \delta \implies \mathbf{E}_{\mu}|\mathrm{L}(h) - \mathrm{L}(h')|^2 \leq \lambda^2 \|h - h'\|^2. \tag{42}$$

We say that L *is weakly λ-smooth at $h \in H$* if L is Gateaux differentiable and the map $h \mapsto \mathrm{L}_r'(h;\cdot)$ is λ-Lipschitz μ-almost surely at h. That is, if for small enough $\delta > 0$ we have

$$\|h - h'\| < \delta \implies \|\mathrm{L}_r'(h;\cdot) - \mathrm{L}_r'(h';\cdot)\| \leq \lambda \|h - h'\|. \tag{43}$$

Note that the norm used here is the operator norm applied to the linear map $\mathrm{L}_r'(h;\cdot) - \mathrm{L}_r'(h';\cdot)$.

3.2.1 Weak convexity of joint composition function

The joint risk function $\mathrm{r}_{\sigma}(\mathrm{L}(h), \theta)$ can be written as a simple composition $(h, \theta) \mapsto (\mathrm{L}(h), \theta) \mapsto \theta + \eta \mathbf{E}_{\mu} \rho_{\sigma}(\mathrm{L}(h) - \theta)$. For any $0 \leq \sigma < \infty$ and any smooth loss, using the preliminary results established Sect. 2.3, it is straightforward to show the weak convexity of this composite function.

Proposition 12 *Let the hypothesis class $\mathcal{H}$ be Banach. Let the loss* L *be locally Lipschitz and weakly λ'-smooth on $\mathcal{H}$. Then, for any* $0 \leq \sigma < \infty$, *defining a σ-dependent factor λ_{σ} as*

$$\lambda_{\sigma} := \begin{cases} 1, & \text{if } \sigma = 0 \\ \pi/(2\sigma), & \text{if } 0 < \sigma < \infty \end{cases}$$

we have that $(h, \theta) \mapsto \mathrm{r}_{\sigma}(\mathrm{L}(h), \theta)$ *is γ-weakly convex with* $\gamma = (1 + \eta\lambda_{\sigma}) \max\{1, \lambda'\}$.

Proof Recall the generic result given in Proposition 8 for the weak convexity of generic composite functions. Our proof here amounts to checking that the assumptions of Proposition 8 are satisfied for the composition $f := \mathrm{r}_{\sigma} \circ f_1$, with $f_1 : \mathcal{H} \times \mathbb{R} \to \mathcal{L}_2(\Omega, \mathcal{F}, \mu) \times \mathbb{R}$ defined as $f_1(h, \theta) := (\mathrm{L}(h), \theta)$.

---

[34] In our particular setting with losses here, the norm used in the numerator of (15) will be the $\mathcal{L}_2$ norm.

To start, let us consider the properties of f_1. Since L is locally Lipschitz and Gateaux differentiable, it follows that L is also Hadamard differentiable.[35] Since the map $h \mapsto \mathrm{L}'_r(h;\cdot) = \mathrm{L}'(h;\cdot)$ is continuous (by weak smoothness), it follows that L is (Fréchet) differentiable.[36] Since $(h, \theta) \mapsto f_1(h, \theta) = (\mathrm{L}(h), \theta)$ just passes θ through the identity, trivially the second component is also differentiable, and the differentiability of both components thus implies f_1 is differentiable.[37] Furthermore, the local Lipschitz property of L is clearly retained by f_1. Evaluating the gradients we have $f_1'(h, \theta)(g, r) = (\mathrm{L}'(h)(g), r)$, and thus using a typical product space norm we have

$$\|f_1'(h_1, \theta_1) - f_1'(h_2, \theta_2)\| = \|\mathrm{L}'(h_1) - \mathrm{L}'(h_2)\| + |\theta_1 - \theta_2|.$$

By weak smoothness of L, it follows that f_1 is $\max\{1, \lambda'\}$-smooth μ-almost surely, where smoothness is in the sense defined in Sect. 2.3.

Next, let us look at properties of r_{σ}. Note that $\rho_{\sigma}(\cdot)$ is trivially 1-Lipschitz in the case of $\sigma = 0$, and for $0 < \sigma < \infty$, the $\pi/2$-Lipschitz property of ρ defined in (3) implies that $\rho_{\sigma}(\cdot)$ is $\pi/(2\sigma)$-Lipschitz. With these basic facts in place, it follows that

$$|\mathrm{r}_{\sigma}(Z_1, \theta_1) - \mathrm{r}_{\sigma}(Z_2, \theta_2)| \leq \left(1 + \eta\lambda_{\sigma}\right)|\theta_1 - \theta_2| + \eta\lambda_{\sigma}\, \mathbf{E}_{\mu}|Z_1 - Z_2|,$$

where the σ-dependent Lipschitz coefficient λ_{σ} is as defined in the statement of the desired result. To obtain bounds in terms of the correct norm, note that

$$\mathbf{E}_{\mu}|Z_1 - Z_2| \leq \mu(\Omega)\sqrt{\mathbf{E}_{\mu}|Z_1 - Z_2|^2} = \sqrt{\mathbf{E}_{\mu}|Z_1 - Z_2|^2},$$

which follows from the fact that μ is a probability, and a simple application of Hölder's inequality.[38] Plugging this into the previous inequality, and noting that it holds for any choice of $Z_1, Z_2 \in \mathcal{L}_2(\Omega, \mathcal{F}, \mu)$ and $\theta_1, \theta_2 \in \mathbb{R}$, it follows that r_{σ} is $(1 + \eta\lambda_{\sigma})$-Lipschitz on $\mathcal{L}_2(\Omega, \mathcal{F}, \mu) \times \mathbb{R}$, and $f_1(\mathcal{H} \times \mathbb{R}) \subset \operatorname{dom} \mathrm{r}_{\sigma}$. Furthermore, from Proposition 3, we have that r_{σ} is convex.

Taking the above points together, if we consider the good event of probability 1 where f_1 satisfies the desired properties, direct application of Proposition 8 to the map $(h, \theta) \mapsto (\mathrm{r}_{\sigma} \circ f_1)(h, \theta)$ yields the desired result. □

Remark 13 The result in the preceding Proposition 12 is rather useful, and it does not require the loss to be convex. When the loss is convex, the analysis becomes somewhat simpler and stronger arguments are naturally possible; composite risks under convex losses and convex, monotonic risk functions is the setting considered by (Ruszczyński and Shapiro 2006, Sec. 3.2), for example.

We have established conditions under which the intermediate joint objective $\mathrm{r}_{\sigma}(\mathrm{L}(h), \theta)$ is weakly convex, and characterized this weak convexity with respect to properties of the underlying risk function and data distribution. Since the data distribution μ is unknown, we can never actually compute $\mathrm{r}_{\sigma}(\mathrm{L}(h), \theta)$. Any learning algorithm will only have access to feedback of a stochastic nature which provides incomplete, noisy information. Our next task is to establish conditions under which the feedback available to the learner is "good enough" to ensure reasonable performance guarantees.

---

[35] (Penot 2012, Prop. 2.25).

[36] (Penot 2012, Prop. 2.51).

[37] (Penot 2012, Prop. 2.52).

[38] (Ash and Doléans-Dade 2000, Sec. 2.4).

3.2.2 Unbiased stochastic feedback

In considering stochastic feedback, recall that $\mathrm{r}_{\sigma}(\mathrm{L}(h), \theta) = \mathbf{E}_{\mu}(f_2 \circ F_1)(h, \theta)$, with F_1 and f_2 given by (7) in the main text. For each $h \in \mathcal{H}$ and $\theta \in \mathbb{R}$, the value $F_1(h, \theta)$ returned by F_1 is a random vector. We shall assume that for any $h \in \mathcal{H}$, the learner can obtain independent random samples of the loss $\mathrm{L}(h)$ and the associated gradient $\mathrm{L}'(h)$. Since $F_1'(h, \theta) = (\mathrm{L}'(h), 1)$ by which we mean $F_1'(h, \theta)(g, r) = (\mathrm{L}'(h)(g), r)$ for all $g \in \mathcal{H}$ and $r \in \mathbb{R}$, clearly the learner can also independently sample from $F_1(h, \theta)$ and $F_1'(h, \theta)$. Sub-differentiability is already guaranteed by Proposition 3, and since ρ_{σ} and η are by design known to the learner, they can readily acquire an element from $\partial f_2(u, \theta)$. Thus if $(\mathrm{L}(h), \theta) \in \operatorname{int}(\operatorname{dom} f_2) = \mathbb{R}^2$ μ-almost surely, it follows that the learner can sample from $\partial f_2(\mathrm{L}(h), \theta) \circ F_1'(h, \theta)$. This is the stochastic *feedback* available to the learner, and when we ask that it be "good enough," this means we require it to be an unbiased estimator of the (Clarke) sub-differential of r_{σ}. The following result gives mild conditions under which this is achieved.

Proposition 14 *Under the conditions of Proposition 12, for any $h \in \mathcal{H}$ and $\theta \in \mathbb{R}$, as long as $\mathbf{E}_{\mu}|\mathrm{L}'(h;\cdot)| < \infty$, the stochastic sub-differential is an unbiased estimator in that*

$$\mathbf{E}_{\mu}\left[\partial f_2(\mathrm{L}(h), \theta) \circ F_1'(h, \theta)\right] \subset \partial_C \mathrm{r}_{\sigma}(\mathrm{L}(h), \theta),$$

and this holds for any choice of $0 \leq \sigma < \infty$.

Proof Using the weak smoothness of L, with probability 1, the map $(h, \theta) \mapsto F_1'(h, \theta)$ is continuous, plus F_1 is locally Lipschitz and $\partial_C F_1(h, \theta) = \{F_1'(h, \theta)\}$.[39] Furthermore, since $F_1'(h, \theta)(g, r) = (\mathrm{L}'(h)(g), r)$, the linearity of $\mathrm{L}'(h)(\cdot)$ implies that $F_1'(h, \theta)(\mathcal{H} \times \mathbb{R}) = \mathbb{R}^2$ μ-almost surely. Since F_1 and f_2 are (locally) Lipschitz, the facts we have just laid out imply a strong chain rule.[40] That is, it holds μ-almost surely that

$$\begin{aligned} \partial_C(f_2 \circ F_1)(h, \theta) &= \partial_C f_2(\mathrm{L}(h), \theta) \circ F_1'(h, \theta) \\ &= \partial f_2(\mathrm{L}(h), \theta) \circ F_1'(h, \theta), \end{aligned} \tag{44}$$

where the second equality follows from the convexity of f_2.

Next, the Lipschitz property of ρ_{σ} implies that f_2 is λ-Lipschitz for some $\lambda > 0$; the actual value is not important for this proof. Using the Lipschitz continuity of f_2 and the fact that $\partial_C F_1(h, \theta) = \{F_1'(h, \theta)\}$ implies $\partial_C \mathrm{L}(h) = \{\mathrm{L}'(h)\}$, we have

---

[39] (Penot 2012, Prop. 5.6).

[40] (Penot 2012, Prop. 5.13).

$$\begin{aligned}
|(f_2 \circ F_1)'_C(h,\theta)(g,r)| &\leq \limsup_{(\alpha,h',\theta')\to(0_+,h,\theta)} \frac{|(f_2 \circ F_1)(h'+\alpha g, \theta'+\alpha r) - (f_2 \circ F_1)(h',\theta')|}{\alpha} \\
&\leq \lambda\left(\limsup_{(\alpha,h')\to(0_+,h)} \frac{|\operatorname{L}(h'+\alpha g) - \operatorname{L}(h')|}{\alpha} + |r| \right) \\
&= \lambda\big(|\operatorname{L}'(h)(g)| + |r|\big).
\end{aligned}$$

Thus, using $\operatorname{r}_\sigma(\operatorname{L}(h),\theta) = \mathbf{E}_\mu(f_2 \circ F_1)(h,\theta)$ and $\mathbf{E}_\mu|\operatorname{L}'(h)(g)| < \infty$ for all $h, g \in \mathcal{H}$, we have

$$\mathbf{E}_\mu\big(f_2 \circ F_1\big)'_C(h,\theta;\cdot) = (\operatorname{r}_\sigma)'_C(\operatorname{L}(h),\theta;\cdot) \tag{45}$$

by a direct application of dominated convergence.[41]

To conclude, taking $G(h,\theta) \in \partial f_2(\operatorname{L}(h),\theta) \circ F_1'(h,\theta)$, by (44) it follows that we have $G(h,\theta) \in \partial_C(f_2 \circ F_1)(h,\theta)$, and thus by definition of the Clarke sub-differential, monotonicity of the integral, and finally (45), we obtain

$$\mathbf{E}_\mu G(h,\theta)(\cdot) \leq \mathbf{E}_\mu(f_2 \circ F_1)'_C(h,\theta;\cdot) = (\operatorname{r}_\sigma)'_C(\operatorname{L}(h),\theta;\cdot). \tag{46}$$

Linearity of $\mathbf{E}_\mu G(h,\theta)(\cdot)$ follows from the linearity of both $G(h,\theta)(\cdot)$ and the integral. Finally, applying (46) we have

$$\sup_{\|(u,r)\|=1} \mathbf{E}_\mu G(h,\theta)(g,r) \leq \sup_{\|(g,r)\|=1} (\operatorname{r}_\sigma)'_C(\operatorname{L}(h),\theta;u,r) < \infty,$$

where finiteness holds because r_σ is locally Lipschitz.[42] Thus $\mathbf{E}_\mu G(h,\theta) \in (\mathcal{H} \times \mathbb{R})^*$, and with (46) we have $\mathbf{E}_\mu G(h,\theta) \in \partial_C \operatorname{r}_\sigma(\operatorname{L}(h),\theta)$ as desired. ☐

Remark 15 The validity of interchanging the operations of (sub-)differentiation and expectation is a topic of fundamental importance in stochastic optimization and statistical learning theory. A useful, modern reference on this topic is included in (Ruszczyński and Shapiro 2003, Ch. 2). A classical reference is Rockafellar and Wets (1982); see also Rockafellar (1968) for a look at measurability of convex integrands. The interchangeability problem appears in various places in the literature over the years, see for example Shapiro (1994) as well as (Kall and Mayer 2005, Ch. 3, Rmk. 2.2). See also (Ruszczyński and Shapiro 2006, Eqn. (3.9)), who refer to generalized versions of a classic result due to Strassen (1965).

3.2.3 Proximity to a nearly-stationary point

Proof of Theorem 5 With all the results established thus far, this proof has just two simple parts. First, we need to show that the objective function of interest is weakly convex, and that we have access to unbiased estimates of the sub-differential; this is done here. This is done using the critical preparatory results in Propositions 12 and 14. Once this has been established, the remaining part just has us applying recent results from the literature for non-asymptotic control of the envelope gradient norm.

---

[41] (Ash and Doléans-Dade 2000, Thm. 1.6.9).

[42] (Penot 2012, Prop. 5.2(b)).

To begin, the assumptions of Proposition 12 are satisfied by A1, which ensures that r_{σ} is γ-weakly convex for $\gamma = (1+\eta\lambda_{\sigma})\max\{1,\lambda\}$. Furthermore, the μ-integrability assumption on $\mathrm{L}'(h)$ lets us use Proposition 14 to ensure that feedback drawn from (9) is such that $\mathbf{E}_{\mu} G_t \in \partial_C \mathrm{r}_{\sigma}(h_t, \theta_t)$ for all t. Furthermore, using A3 implies that $\mathbf{E}[G_t \mid G_{[t-1]}] \in \partial_C \mathrm{r}_{\sigma}(h_t, \theta_t)$ for all t, since Algorithm 1 uses G_t sampled via (9).

The desired result follows from an application of (Davis and Drusvyatskiy 2019, Thm. 3.1), where their objective function f corresponds to our r_{σ}.[43] While their proof is given for the case of $\mathcal{H} = \mathbb{R}^d$, using assumption A2, if we leverage our characterization of weak convexity (Proposition 7), and replace their Lemma 2.2 with our (32), it is straightforward to see that their insights extend to arbitrary Hilbert spaces using the usual norm induced by the inner product. Thus with the moment bound A4 in hand, the generalized result can be applied to Algorithm 1, for objective function r_{σ}, which has just been proved to be γ-weakly convex. The desired result follows immediately. $\square$

Appendix 4 Helper results

In this section, we provide some standard results that are leveraged in the main paper.

4.1 Useful results based on Lipschitz properties

Let $\mathcal{X}$ be a normed linear space, and let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be convex and λ-Lipschitz. If f is sub-differentiable at a point x, then using the definition of the sub-differential, we have that

$$|\langle x' - x, \partial f(x)\rangle| \leq |f(x') - f(x)| \leq \lambda \|x' - x\|.$$

It immediately follows that

$$\|\partial f(x)\| \leq \lambda. \tag{47}$$

That is, all sub-gradients of f at x have norm no greater than the Lipschitz coefficient λ.

Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be differentiable on $U \subset \mathcal{X}$, an open set. Further, assume that the derivative is λ-Lipschitz on U, that is, for each $x, x' \in U$, we have $\|f'(x) - f'(x')\| \leq \lambda \|x - x'\|$. First-order Taylor approximations have direct analogues in this general setting, as the following result shows.[44]

Proposition 16 *Let $f : \mathcal{X} \to \mathcal{Y}$ be differentiable on an open set $U \subset \mathcal{X}$, with $\mathcal{X}$ and $\mathcal{Y}$ assumed to be Banach. If $f'(\cdot)$ is λ-Lipschitz on U, then for any $x, u \in U$ such that $x + u \in U$, we have*

$$\|f(x+u) - f(x) - f'(x)(u)\| \leq \frac{\lambda}{2}\|u\|^2.$$

---

[43] It also relies on the observation that a *proximal* stochastic sub-gradient update using the indicator function of C as a regularizer is equivalent to the *projected* sub-gradient update we do here.

[44] See for example (Luenberger 1969, Sec. 7.3, Prop. 2–3) and (Nesterov 2004, Ch. 1).

4.2 Radial derivatives of convex functions

Say a function $f : \mathcal{V} \to \overline{\mathbb{R}}$ is convex. Take any $u, v \in \operatorname{dom} f$, and any scalar $c \geq 0$ such that $v + c(v - u) \in \operatorname{dom} f$. Then writing $u' := v + c(v - u)$, note that

$$v = \frac{1}{1+c}\left(u' + cu\right) = (1-\beta)u' + \beta u,$$

where $\beta := c/(1+c) \in [0,1)$. By convexity we have $f((1-\beta)u' + \beta u) \leq (1-\beta)f(u') + \beta f(u)$. Filling in definitions and rearranging we have

$$f(v + c(v-u)) - f(v) \geq c(f(v) - f(u)). \tag{48}$$

Note that this can be done for any pair of u, v and scalar c that keeps the relevant points on the domain. Clearly this property is necessary for convexity, but it is in fact also sufficient.[45]

For any function $f : \mathcal{V} \to \overline{\mathbb{R}}$ and open set $U \subset \mathcal{X}$, fix a point $x \in U$. We denote the difference quotient of f at x, incremented in the direction u, modulated by scalar $\alpha \neq 0$ as

$$q(\alpha) := q(\alpha; f, x, u) := \frac{f(x + \alpha u) - f(x)}{\alpha}. \tag{49}$$

Consider the map $g(t) := f(x + tu) - f(x)$, with all elements but $t \geq 0$ fixed. When f is convex, direct inspection immediately shows that $t \mapsto g(t)$ is convex. For any $0 \leq t_1 < t_2$, take some $t' \in (t_1, t_2)$. Clearly, there exists a $\beta \in (0,1)$ such that $t' = \beta t_1 + (1-\beta)t_2$. Then, we have

$$\begin{aligned}\frac{g(t') - g(t_1)}{t' - t_1} = \frac{g(\beta t_1 + (1-\beta)t_2) - g(t_1)}{(1-\beta)(t_2 - t_1)} &\leq \frac{\beta g(t_1) + (1-\beta)g(t_2) - g(t_1)}{(1-\beta)(t_2 - t_1)}\\ &= \frac{g(t_2) - g(t_1)}{t_2 - t_1},\end{aligned}$$

where the inequality follows from convexity of g. If we use this inequality in the special case of $t_1 = 0$, alongside the basic relation $q(\alpha) = (g(\alpha) - g(0))/\alpha$, it immediately follows that $\alpha \mapsto q(\alpha)$ is monotonic (non-decreasing) on the positive reals. Furthermore, the set $\{q(\alpha) : \alpha > 0\}$ is bounded below. To see this, take some $\gamma > 0$ small enough that $x - \gamma u \in \operatorname{dom} f$, and note that by direct application of convexity and the basic property (48), it follows that

$$\left(\frac{\alpha}{\gamma}\right)(f(x) - f(x - \gamma u)) \leq f\left(x + \frac{\alpha}{\gamma}(x - (x - \gamma u))\right) - f(x) = f(x + \alpha u) - f(x).$$

That is, dividing both sides by α, we have

$$\left(\frac{1}{\gamma}\right)(f(x) - f(x - \gamma u)) \leq \frac{f(x + \alpha u) - f(x)}{\alpha} = q(\alpha; f, u). \tag{50}$$

Since the choice of $\gamma > 0$ depends only on x and u, and is free of α, it follows that the set $\{q(\alpha) : \alpha > 0\}$ is bounded below, as desired. Using this boundedness alongside the monotonicity of $\alpha \mapsto q(\alpha)$, we have that the infimum is finite. Thus, recalling the definition (13) of the radial derivative of f at x in the direction u, since we have

---

[45] For example, see (Nesterov 2004, Thm. 3.1.1).

$$f_r'(x;u) = \lim_{\alpha \to 0_+} \frac{f(x+\alpha u) - f(x)}{\alpha} = \inf\{q(\alpha;f,x,u) : \alpha > 0\}, \tag{51}$$

it follows immediately that the radial derivative always exists (i.e., $f'(x;u) \in \mathbb{R}$). Note also that using convexity, direct inspection shows that for all u we have

$$f(u) - f(x) \geq f'(x;u-x). \tag{52}$$

Furthermore, it is easily verified that whenever $x \in \operatorname{dom} f$, the map $u \mapsto f'(x;u)$ is sub-additive and positively homogeneous, i.e., a sub-linear functional.[46] The basic facts of interest here are summarized in the following proposition.

Proposition 17 (Difference quotients for convex functions) *Let $\mathcal{V}$ be a vector space. If function $f : \mathcal{V} \to \overline{\mathbb{R}}$ is proper and convex, then it is radially differentiable on* int (dom$f$).

Proof The desired result follows immediately from previous discussion leading up to (51), and the fact that if x is an interior point of the effective domain of f, it follows that for any $u \in \mathcal{V}$, we can find a $\gamma > 0$ small enough that $x - \gamma u \in \operatorname{dom} f$, which means we can apply the lower bound of (50) to the difference quotients $q(\alpha;f,x,u)$ indexed by $\alpha > 0$. ☐

Appendix 5 Loss example

Example 18 While stated with a somewhat high degree of abstraction, let us give a concrete example to emphasize that the assumptions of Proposition 12 are readily satisfied under natural and important learning settings. Consider the regression problem, where we observe random pairs $(X, Y) \sim \mu$, assuming that X is a finite-dimensional real-valued random vector, and Y is a real-valued random variable, related to the inputs by the relation $Y = h^*(X) + \epsilon$, where ϵ is a zero-mean random noise term. For simplicity, let h^* be a continuous linear map, and let $\mathcal{H}$ be the set of all continuous linear maps on the space that X is distributed over. Finally, let the loss by the squared error, such that

$$\begin{aligned} \mathrm{L}(h) &= (h(X) - Y)^2 = (\langle X, h - h^* \rangle - \epsilon)^2 \\ \mathrm{L}'(h)(u) &= 2(\langle X, h - h^* \rangle - \epsilon)\langle u, X \rangle. \end{aligned}$$

Since we make almost no assumptions on the nature of the underlying noise distribution, clearly both the losses and the “gradients” can be unbounded and heavy-tailed. Fix any $h_0 \in \mathcal{H}$, and note that for any $h \in \mathcal{H}$, we have

$$\begin{aligned} \mathrm{L}(h) - \mathrm{L}(h_0) &= \langle X, h - h^* \rangle^2 - \langle X, h_0 - h^* \rangle^2 - 2\epsilon\langle X, h - h_0 \rangle \\ &= \langle X, h - h_0 \rangle \langle X, (h - h^*) + (h_0 - h^*) \rangle - 2\epsilon \langle X, h - h_0 \rangle. \end{aligned}$$

---

[46] This means that the Hahn-Banach theorem can be applied to construct a linear functional g bounded above as $g(u) \leq f'(x;u)$, for all u. See for example (Luenberger 1969, Sec. 5.4) or (Ash and Doléans-Dade 2000, Thm. 3.4.2). This g is not necessarily a sub-gradient of f at x, since it need not be continuous in general; such functions are sometimes called *algebraic sub-gradients* (Ruszczyński and Shapiro 2006, Sec. 3).

Table 1 A summary of the benchmark datasets used for performance evaluation

Dataset	Size	Input features	Number of classes	Model dimension
`adult`	45,222	105	2	210
`australian`	690	43	2	86
`cifar10`	60,000	3072	10	30,720
`cod_rna`	331,152	8	2	16
`covtype`	581,012	54	7	378
`emnist_balanced`	131,600	784	47	36,848
`fashion_mnist`	70,000	784	10	7840
`mnist`	70,000	784	10	7840

Absolute values can be bounded above as

$$|\mathrm{L}(h) - \mathrm{L}(h_0)| \leq \|X\|^2 \|h - h_0\| \left(\|h - h^*\| + \|h_0 - h^*\| \right) + 2|\epsilon| \|X\| \|h - h_0\|.$$

It follows immediately that as long as $\mathbf{E}_{\mu}\|X\|^4 < \infty$, we have that the local Lipschitz property (42) of the loss is satisfied, for arbitrary choice of h_0.

As for the weak smoothness requirement on the loss, note that

$$\begin{aligned} \|\mathrm{L}'(h) - \mathrm{L}'(h_0)\| &= \sup_{\|u\|=1} \langle u, \mathrm{L}'(h) - \mathrm{L}'(h_0) \rangle \\ &= \sup_{\|u\|=1} 2(\langle X, h - h_0 \rangle)\langle u, X \rangle \\ &\leq 2\|X\|^2 \|h - h_0\|. \end{aligned}$$

Thus, if the random inputs X are μ-almost surely bounded, the desired smoothness condition (43) holds. Note that this does not preclude heavy-tailed losses and gradients since no additional assumptions have been made on the noise term.

Appendix 6 Empirical supplement

Due to limited space, we could only include key details and a few representative results in Sect. 4 of the main text. Here we fill in those additional details. To begin, all our numerical experiments have been implemented entirely in Python (v. 3.8) using the following additional open-source software: Jupyter notebook (for interactive demos),[47] matplotlib (v. 3.4.1, for all visuals),[48] PyTables (v. 3.6.1, for dataset handling),[49] NumPy (v. 1.20.0, for almost all computations),[50] and SciPy (v. 1.6.2, for random variable statistics and special functions). In the following paragraphs, we provide information about the benchmark datasets used, as well as several figures including additional experimental results.

---

[47] https://jupyter.org/

[48] https://matplotlib.org/

[49] https://www.pytables.org/

[50] https://numpy.org/

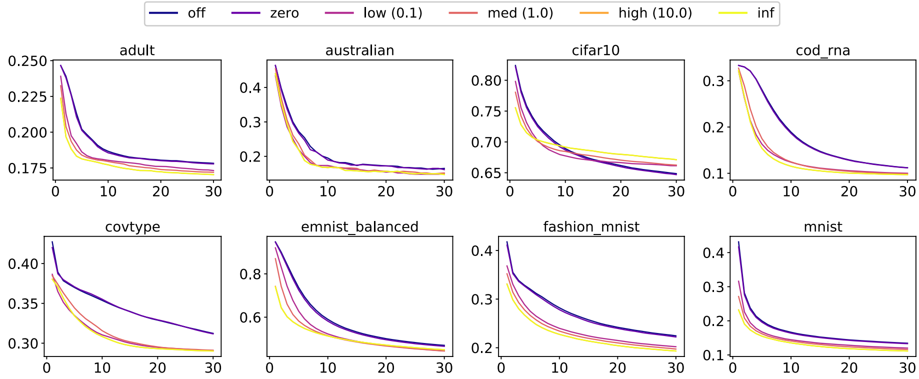


Fig. 6 Additional average classification error trajectories over epochs

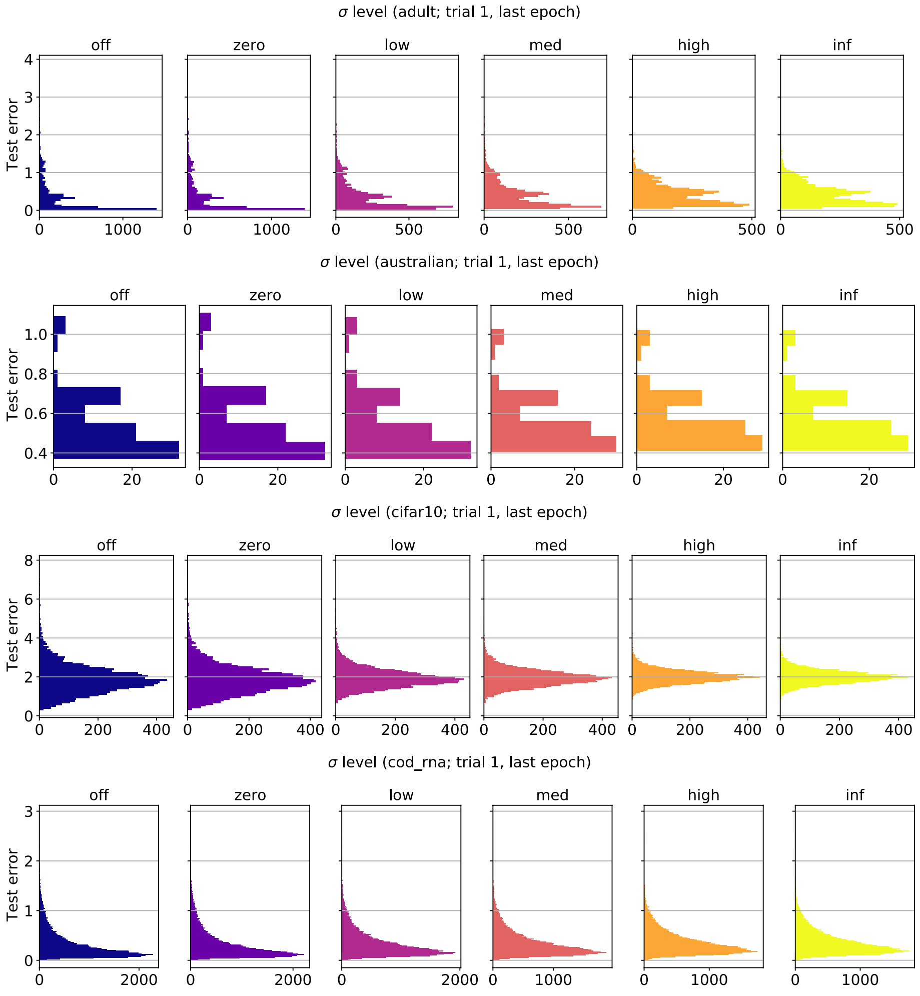


Fig. 7 Additional test error histograms

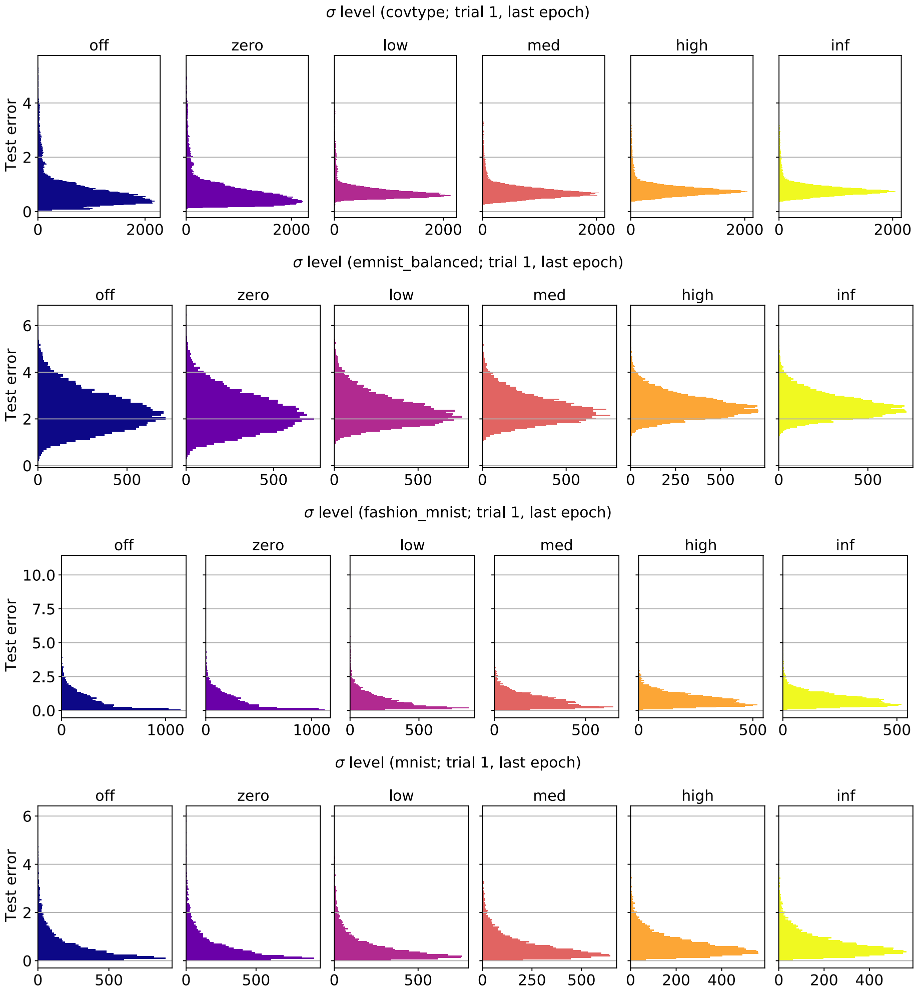


Fig. 8 Additional test error histograms

Dataset description

The real-world benchmark datasets used in our classification tests are as follows: `adult`,[51] `australian`,[52] `cifar10`,[53] `cod_rna`,[54] `covtype`,[55]

---

[51] https://archive.ics.uci.edu/ml/datasets/Adult

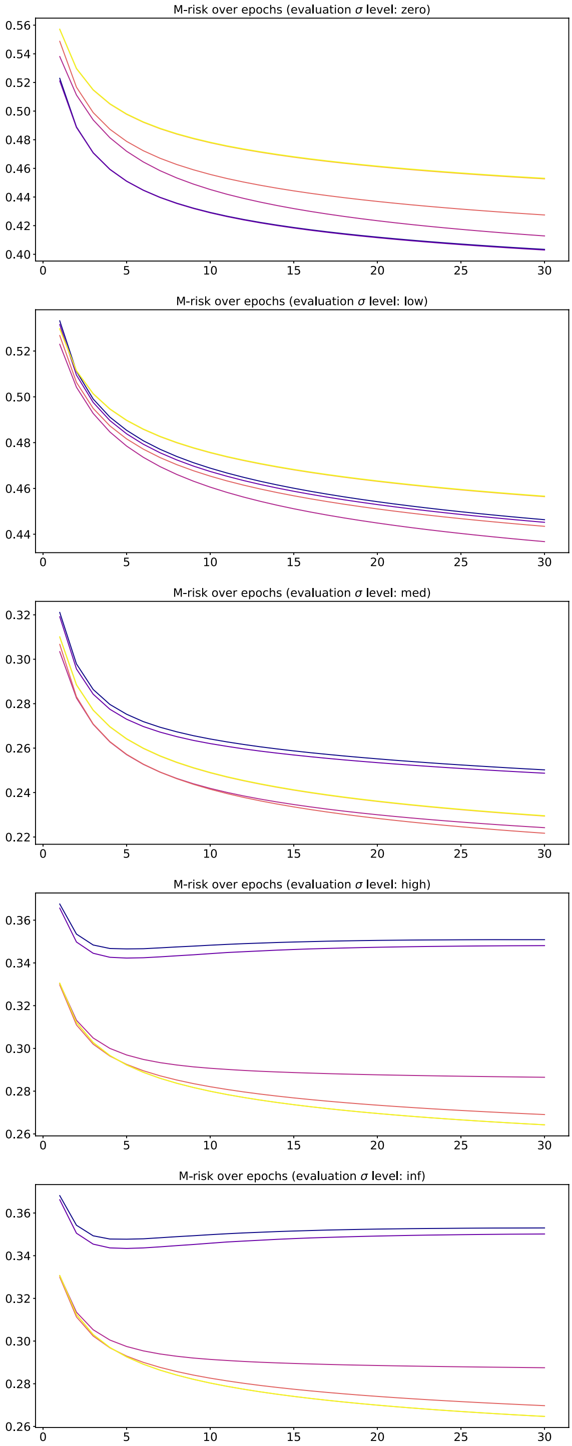
[52] https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval)

[53] https://www.cs.toronto.edu/~kriz/cifar.html

[54] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

[55] https://archive.ics.uci.edu/ml/datasets/covertype

Fig. 9 Empirical mean estimates of R_{σ} for a variety of $\sigma \in [0, \infty]$ settings. The colored curves correspond to different σ settings in running Algorithm 1, just as in previous plots, whereas the distinct plots correspond to the different σ used in *evaluation* (Color figure online)



`emnist_balanced`,[56] `fashion_mnist`,[57] and `mnist`.[58] See Table 1 for a summary. Further background on all datasets is available at the URLs provided in the footnotes. Dataset size reflects the size after removal of instances with missing values, where applicable. For all datasets with categorical features, the "input features" given in Table 1 represents the number of features after doing a one-hot encoding of all such features. The "model dimension" is just the product of the number of input features and the number of classes, since we are using multi-class logistic regression (one linear model for each class). All categorical features are given a one-hot representation. All input features are standardized to take values on the unit interval [0, 1]. As mentioned in the main text, we have prepared a GitHub repository that includes code for both re-creating the empirical tests and preprocessing the data (https://github.com/feedbackward/mrisk).

Additional results

In Figs. 6, 7 and 8, we give additional results that complement Fig. 4 in the main text. The trends in terms of the histograms of test loss distributions are essentially uniform across this wide variety of datasets. We also see that a sharply-concentrated logistic (test) loss tends to correlate with better classification error (average zero-one error), with `cifar10` being the only exception to this trend. As another point not raised in the main text, intuitively we would hope that Algorithm 1 performs well in terms of the risk R_{σ} corresponding to its particular σ setting; we have found this to be true across the benchmark datasets studied here. See Fig. 9 for an example from the `adult` dataset. Moving from top to bottom, the order of colors shows a rather clear reversal. Very similar trends can be observed on the other datasets as well. In estimating R_{σ} on the test set, we use an empirical mean estimate of r_{σ}, and then minimize with respect to θ using the `minimize_scalar` function of the SciPy (v. 1.6.2) `optimize` module.

Author Contributions MJH was responsible for all elements of this work.

Funding This work was supported by JST ACT-X Grant Number JPMJAX200O, Japan.

Availability of data and materials Not applicable.

Code availability A GitHub repository including all code and interactive demos is available online. (https://github.com/feedbackward/mrisk) By using the code provided and following the documentation, all of the experimental results, as well as the figures presented here can be re-created on demand.

Declarations

Competing interests The author declares no competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

---

[56] https://www.nist.gov/itl/products-and-services/emnist-dataset

[57] https://github.com/zalandoresearch/fashion-mnist

[58] http://yann.lecun.com/exdb/mnist/

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

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

Ash, R. B., & Doléans-Dade, C. A. (2000). *Probability and measure theory* (2nd ed.). NP: Academic Press.

Barbu, V., & Precupanu, T. (2012). *Convexity and optimization in banach spaces* (4th ed.). Berlin: Springer Science & Business Media.

Bauschke, H. H., Borwein, J. M., & Combettes, P. L. (2003). Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization, 42*(2), 596–636.

Bauschke, H. H., & Combettes, P. L. (2017). *Convex analysis and monotone operator theory in Hilbert spaces* (2nd ed.). Berlin: Springer.

Bertsekas, D. P. (2015). *Convex optimization algorithms*. Nashua: Athena Scientific.

Bhat, S.P., & Prashanth, L.A. (2020). Concentration of risk measures: A Wasserstein distance approach. Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

Bottou, L., Curtis, F.E., Nocedal, J. (2016). Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation, 11*(7), 1493–1517.

Brownlees, C., Joly, E., & Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics, 43*(6), 2507–2536.

Daniilidis, A., & Malick, J. (2005). Filling the gap between lower-C^1 and lower-C^2 functions. *Journal of Convex Analysis, 12*(2), 315–329.

Davis, D., & Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization, 29*(1), 207–239.

Drusvyatskiy, D., & Paquette, C. (2019). Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming, 178*, 503–558.

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

Duchi, J. C., & Ruan, F. (2018). Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization, 28*(4), 3229–3259.

Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization, 23*(4), 2341–2368.

Hashimoto, T.B., Srivastava, M., Namkoong, H., Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)* Vol. 80, pp. 1929–1938.

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

Holland, M.J., & Haress, E.M. (2021a). Learning with risk-averse feedback under potentially heavy tails. In *24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)* Vol. 130.

Holland, M.J., & Haress, E.M. (2021b). Spectral risk-based learning using unbounded losses. arXiv preprint arXiv:2105.04816.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*(1), 73–101.

Johnson, R., & Zhang, T. (2014). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems* 26 (NIPS 2013) pp. 315–323.

Jourani, A., Thibault, L., & Zagrodny, D. (2014). Differential properties of the Moreau envelope. *Journal of Functional Analysis, 266*(3), 1185–1237.

Kall, P., & Mayer, J. (2005). Stochastic linear programming. *International Series in Operations Research and Management Science*.

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

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

Leqi, L., Prasad, A., Ravikumar, P.K. (2019). On human-aligned risk minimization. *Advances in Neural Information Processing Systems 32* (NeurIPS 2019).

Le Roux, N., Schmidt, M., Bach, F.R. (2013). A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems 25* (NIPS 2012) pp. 2663–2671.

Luenberger, D. G. (1969). *Optimization by vector space methods*. New Jersey: Wiley.

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

Maurer, A., & Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory* (COLT).

Namkoong, H., & Duchi, J.C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems* 29 (NIPS 2016) Vol. 29, pp. 2208–2216.

Nemirovsky, A. S., & Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. New Jersey: Wiley.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*. Berlin: Springer.

Penot, J. P. (2012). *Calculus without derivatives*. Berlin: Springer.

Poliquin, R. A., & Rockafellar, R. T. (1996). Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society, 348*(5), 1805–1838.

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

Reyzin, L., & Schapire, R.E. (2006). How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML 2006) (pp. 753–760).

Rockafellar, R. T. (1968). Integrals which are convex functionals. *Pacific Journal of Mathematics, 24*(3), 525–539.

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

Rockafellar, R. T., & Wets, R. J. B. (1982). On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics: An International Journal of Probability and Stochastic Processes, 7*(3), 173–182.

Rockafellar, R. T., & Wets, R. J. B. (1998). *Variational Analysis*. Berlin: Springer.

Ruszczyński, A., & Shapiro, A. (2003). Stochastic programming models handbooks in operations research and management. *Science, 10*, 1–64.

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

Ruszczyński, A., & Shapiro, A. (2006). Optimization of risk measures. *Probabilistic and randomized methods for design under uncertainty* (pp. 119–157). Berlin: Springer.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press.

Shalev-Shwartz, S., & Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research, 14*, 567–599.

Shamir, O., & Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning* pp. 71–79.

Shapiro, A. (1994). Quantitative stability in stochastic programming. *Mathematical Programming, 67*(1–3), 99–108.

Soueycatt, M., Mohammad, Y., & Hamwi, Y. (2020). Regularization in banach spaces with respect to the bregman distance. *Journal of Optimization Theory and Applications, 185*, 327–342.

Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics, 36*(2), 423–439.

Vapnik, V. N. (1999). *The nature of statistical learning theory* (2nd ed.). Berlin: Springer.

Williamson, R.C., & Menon, A.K. (2019). Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning* (ICML) pp. 6786–6797.

Zhai, R., Dan, C., Kolter, J.Z., Ravikumar, P. (2021). DORO: Distributional and outlier robust optimization. In *38th International Conference on Machine Learning* (ICML) Vol. 139, pp. 12345–12355.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.