Criterion Collapse and Loss Distribution Control

Matthew J. Holland [1]

Abstract

In this work, we consider the notion of "criterion collapse," in which optimization of one metric implies optimality in another, with a particular focus on conditions for collapse into error probability minimizers under a wide variety of learning criteria, ranging from DRO and OCE risks (CVaR, tilted ERM) to non-monotonic criteria underlying recent ascent-descent algorithms explored in the literature (Flooding, SoftAD). We show how collapse in the context of losses with a Bernoulli distribution goes far beyond existing results for CVaR and DRO, then expand our scope to include surrogate losses, showing conditions where monotonic criteria such as tilted ERM cannot avoid collapse, whereas non-monotonic alternatives can.

1. Introduction

As machine learning systems become more widespread and integrated in our daily lives, the tension between performance metrics we ideally wish to optimize (at test time) and those we actually optimize in practice (at training time) becomes increasingly important. Broadly speaking, tasks are characterized by *losses* and what we will call *learning criteria*. Losses are diverse, depending on the underlying problem to be solved (e.g., classification, regression, ranking, clustering, etc.) and a wide range of application-specific needs. Losses depend on random data, and are themselves random. In principle, performance can be described in terms of random loss distributions, but working with complex and unwieldy distributions is not practical; a concise numerical summary is usually desired both for designing objectives to optimize, as well as for disseminating interpretable results. Providing such a summary is the role of learning criteria. In contrast with the diversity of losses, there is a single *de facto* standard learning criterion, namely the expected value of the loss distribution, a criterion which is central to the "general setting of the learning problem" (Vapnik, 1999) that is pervasive in both theory and practice.

In recent years, however, the need to capture a wider variety of performance metrics at training time has motivated the use of novel learning criteria going beyond the expected value. Arguably the best-known families of such criteria are those which control sensitivity to the right tail of the loss distribution, e.g., CVaR (Curi et al., 2020), positive-tilted ERM (Li et al., 2021), and Cressie-Read DRO (Duchi & Namkoong, 2021), though criteria designed to capture both tails at once have also begun to appear (Holland, 2022). While all of these criteria can be clearly shown to be distinct from the expected value, when such criteria are optimized at training time, can we expect the outcome to actually change? This question is somewhat delicate, since it depends on the nature of the loss distribution. For example, when we are working with the zero-one loss for classification, or any binary performance indicator for that matter, previous work has noted that vanilla empirical risk minimization (ERM) is always sufficient for minimizing DRO and CVaR (Hu et al., 2018; Zhai et al., 2021), making it meaningless to distinguish between optimizing such criteria.

What about when we make use of surrogate loss functions for training? The traditional theory of transferring consistency from a surrogate to a target loss is tied tightly to the expected value criterion (and its linearity) (Bartlett et al., 2006; Reid & Williamson, 2011), meaning that a change in criterion can drastically change the relations between loss distributions. For example, given a model which is CVaR-optimal in terms of a surrogate loss, it may or may not be CVaR-optimal in terms of the zero-one loss; understanding when such properties hold is critical to ensure the learning algorithms we use in practice are aligned with our true objectives. For many classes of learning criteria, however, such questions remain completely unexplored.

In this work, we consider the notion of "criterion collapse," in which optimization of one metric implies optimality in another, with a particular focus on conditions for collapse into error probability minimizers (i.e., minimizers of the expected zero-one loss) under a wide variety of learning criteria, including learning scenarios that involve a surrogate loss. We start in §2 by showing how collapse in the context of losses with a Bernoulli distribution goes far beyond CVaR

[1]SANKEN, Osaka University, Japan. Correspondence to: Matthew J. Holland <matthew-h@ar.sanken.osaka-u.ac.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

and DRO. We then expand our scope to include surrogate losses in §3. First, we highlight cases where optimality diverges across loss distributions under a common criterion (§3.1), and then on the other hand show conditions where for *monotonic* criteria (e.g., positive-tilted ERM), collapse into error probability minimizers is unavoidable (§3.2), though the use of non-monotonic alternatives can be used to avoid this when such collapse is not desired (§3.3). The main limitation of our results in §3 is that we only cover margin-type losses for binary classification. Extensions to asymmetric, multi-class surrogate losses is left as future work. We complement our basic theory with a set of experiments in §4, training non-linear neural network models (e.g., ResNet-34) for image classification from scratch, comparing across a variety of learning criteria with a common base loss. We find that when criterion selection is optimized for validation accuracy, the non-monotonic criteria (with clear links to ascent-descent learning algorithms) provide an appealing balance across other metrics such as model norm and average (surrogate) loss at test time.

2. Criterion Collapse

Underlying this section is the following simple question: *is it meaningful to introduce a new criterion?* Even if two criteria are distinct in a mathematical sense, i.e., they return different values for the same loss distribution, if the solution set of one criterion completely includes the other, then an obvious *sufficiency* property holds; no tradeoffs between criteria arise. When we have a sufficient condition (for optimality in both criteria) that is easy to satisfy, it essentially renders one criterion meaningless. Here we will look at how such a phenomenon occurs frequently under binary distributions.

2.1. Preliminaries

Basic notation Let $\mathcal{X} \times \mathcal{Y}$ denote our underlying data space. Random data points will be denoted by individual random pairs distributed on $\mathcal{X} \times \mathcal{Y}$. For a single data point, often to represent "data at test time," we write (X, Y). When we are dealing with non-random quantities, X and Y will be replaced by x and y respectively. As general-purpose notation for decisions, we write h, assumed to be an element of a set of admissible decisions $\mathcal{H}$ (the "hypothesis class"). Throughout this paper, when we use $\mathbf{E}$ and $\mathbf{P}$ to represent expectation and probability, it will always be with respect to the distribution of (X, Y), unless otherwise noted. For an arbitrary real-valued function $f : \mathcal{H} \to \mathbb{R}$, we denote the set of minimizers by $\arg\min_{h \in \mathcal{H}} f(h) \subset \mathcal{H}$, with $\arg\min_{h \in \mathcal{H}} f(h) = \emptyset$ in the case that no minimizers exist in $\mathcal{H}$. We remark that unlike h (always $h : \mathcal{X} \to \mathcal{Y}$), the notation f is not reserved here. We use f and also g to denote various different helper functions throughout the paper.

Loss function and criterion mapping Central to this paper is the notion of a numerical *loss* that can be used to provide feedback to a learning algorithm, or as a metric for evaluation after learning is complete. A *loss function* $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ maps (decision, data point) pairs to real values $\ell(h; x, y)$, called losses. When the data is random, so is the loss. We denote random losses by $\mathsf{L}(h) := \ell(h; \mathsf{X}, \mathsf{Y})$, making the dependence on decision $h \in \mathcal{H}$ explicit in our notation. Running over the set $\mathcal{H}$, we end up with a set of random losses, denoted by $\mathcal{L} := \{\mathsf{L}(h) : h \in \mathcal{H}\}$. We will sometimes just refer to individual random losses $\mathsf{L} \in \mathcal{L}$ without specifying which $h \in \mathcal{H}$ is associated with L. The other basic notion we require is that of a *criterion mapping*, denoted $\mathrm{C} : \mathcal{L} \to \mathbb{R}$, which maps random losses $\mathsf{L} \in \mathcal{L}$ to numerical values $\mathrm{C}(\mathsf{L})$, called criteria. As mentioned in §1, the traditional criterion is $\mathrm{C}(\mathsf{L}) = \mathbf{E}[\mathsf{L}]$, often called the "risk," but we will go well beyond this criterion in the following sub-sections.

2.2. Random error and criterion collapse

The most basic and pervasive loss function used for evaluation is the "zero-one loss," a simple classification error penalty that we denote as

$$\ell_{01}(h; x, y) := \begin{cases} 1, & \text{if } h(x) \neq y \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Denoting the random error by $\mathsf{L}_{01}(h) := \ell_{01}(h; \mathsf{X}, \mathsf{Y})$, it follows a Bernoulli distribution that is completely determined by its expected value, the error probability $\mathbf{E}[\mathsf{L}_{01}(h)] = \mathbf{P}\{h(\mathsf{X}) \neq \mathsf{Y}\}$. Since this quantity will appear frequently, we reserve the following symbols for the error probability and its solution set:

$$\mathcal{E}(h) := \mathbf{P}\{h(\mathsf{X}) \neq \mathsf{Y}\}, \qquad \mathcal{H}^*_{\mathcal{E}} := \arg\min_{h \in \mathcal{H}} \mathcal{E}(h). \tag{2}$$

The expected value $\mathsf{L} \mapsto \mathbf{E}[\mathsf{L}]$ also happens to be the traditional choice of criterion mapping in machine learning, but as discussed earlier in §1, a variety of new choices for criterion mapping have arisen in the context of "risk-sensitive" learning.

In general, by introducing new criteria (instead of the expected value), one gains an additional degree of freedom (beyond choice of loss function) in terms of what qualities we evaluate, and how we quantify such qualities. However, for loss functions such as ℓ_{01} in (1), the resulting random loss L_{01} has such a simple distribution that many criteria "collapse" into the expected value. Two important special cases of this phenomenon have been noted in the previous literature. We recall these two cases here in chronological order, adapted to our notation for readability.

Theorem 2.1 (DRO criterion; Hu et al. (2018, Thm. 1))**.** *For arbitrary random loss $\mathsf{L} \in \mathcal{L}$, denote the distributionally*

robust optimization (DRO) criterion by

$$\mathrm{DRO}(\mathsf{L}) \coloneqq \sup_{\mu \in \mathcal{P}} \mathbf{E}_{\mu}[\mathsf{L}]$$

where the "uncertainty set" $\mathcal{P}$ is taken to be a ball centered at some pre-defined data distribution on $\mathcal{X} \times \mathcal{Y}$, with finite radius measured by a valid f-divergence. Under zero-one loss $\mathsf{L} = \mathsf{L}_{01}$, error probability minimizers are always optimal in terms of the DRO criterion, namely we have

$$\mathcal{H}^*_{\mathcal{E}} \subset \arg\min_{h \in \mathcal{H}} \mathrm{DRO}(\mathsf{L}_{01}(h))$$

with $\mathcal{H}^*_{\mathcal{E}}$ defined earlier in (2).

Another very closely related insight has been presented in the literature, this time looking at conditional value-at-risk (CVaR), a well-studied criterion that is central to the quantification of financial risk.

Theorem 2.2 (CVaR criterion; Zhai et al. (2021, Prop. 1))**.** *Again taking any loss $\mathsf{L} \in \mathcal{L}$, we write the conditional value-at-risk (CVaR) criterion as*

$$\mathrm{CVaR}(\mathsf{L}) \coloneqq \inf_{\theta \in \mathbb{R}} \left[\theta + \frac{1}{1-\beta} \mathbf{E}[(\mathsf{L} - \theta)_+] \right]$$

where $0 < \beta < 1$ controls the degree of right-tail sensitivity, and $(\cdot)_+ \coloneqq \max\{0, \cdot\}$. Under the zero-one loss $\mathsf{L} = \mathsf{L}_{01}$, error probability minimizers are optimal in terms of CVaR, i.e.,

$$\mathcal{H}^*_{\mathcal{E}} \subset \arg\min_{h \in \mathcal{H}} \mathrm{CVaR}(\mathsf{L}_{01}(h)).$$

The basic message of Theorems 2.1 and 2.2 is clear: under losses with a Bernoulli distribution, it is essentially meaningless to consider DRO and CVaR as distinct from the expected value, since minimizing the expected loss is always sufficient for optimality in terms of DRO and CVaR as well. In §2.3 to follow, we show that this "collapse" into error probability minimizers arises for criteria going well beyond DRO and CVaR.

2.3. Which classes lead to collapse?

Here we take a look at several large classes of criteria, showing how most collapse in the sense illustrated in §2.2.

2.3.1. Expectation of fixed function

Starting with the simplest class, we consider criteria that are computed as

$$\mathsf{L} \mapsto \mathbf{E}[f(\mathsf{L})] \tag{3}$$

where $f : \mathbb{R} \to \mathbb{R}$ is any function such that the expectation is finite. By setting $\mathsf{L} = \mathsf{L}_{01}$ and reflecting dependence on $h \in \mathcal{H}$, note that

$$\mathbf{E}[f(\mathsf{L}_{01}(h))] = f(0) + \mathcal{E}(h)\,(f(1) - f(0))$$

where $\mathcal{E}(\cdot)$ is the error probability defined in (2). If $f(0) = f(1)$ happens to hold, the criterion is constant and thus not very interesting. When $f(0) \neq f(1)$, minimizers of $\mathbf{E}[f(\mathsf{L}_{01}(\cdot))]$ are easy to characterize: they either minimize or maximize $\mathcal{E}(\cdot)$, depending on whether $f(1) > f(0)$ or not. As such, no matter how f is designed, the result of minimizing such a criterion will always be one of these two extremes.

2.3.2. Quantiles

As a natural alternative to the mean, let us consider a standard definition of quantiles, namely the β-level "left quantile" of the random loss, defined for $\mathsf{L} \in \mathcal{L}$ by

$$\mathrm{Q}_{\beta}(\mathsf{L}) \coloneqq \min\{x \in \mathbb{R} : \mathbf{P}\{\mathsf{L} \leq x\} \geq \beta\} \tag{4}$$

for all $0 < \beta \leq 1$, and $\mathrm{Q}_{\beta}(\mathsf{L}) \coloneqq -\infty$ for $\beta = 0$. While in general the relation between quantiles and the mean can be very complicated, for the special case of Bernoulli losses, this relation is very simple for all choices of β, and we can readily show how the quantile criterion collapses to the mean.

Proposition 2.3 (Collapse of left quantiles)**.** *Given Q_{β} as defined in (4) and $\mathsf{L} = \mathsf{L}_{01}$, we have*

$$\mathcal{H}^*_{\mathcal{E}} \subset \arg\min_{h \in \mathcal{H}} \mathrm{Q}_{\beta}(\mathsf{L}_{01}(h))$$

for all probability levels $0 \leq \beta \leq 1$.

Remark 2.4 (Related case: right quantiles)**.** Another very similar discussion can be carried out with the "right quantiles" defined by

$$\mathrm{Q}^-_{\beta}(\mathsf{L}) \coloneqq \max\{x \in \mathbb{R} : \mathbf{P}\{\mathsf{L} < x\} \leq \beta\} \tag{5}$$

for $0 \leq \beta < 1$, and $\mathrm{Q}^-_{\beta}(\mathsf{L}) \coloneqq \infty$ for $\beta = 1$.

2.3.3. Distribution dependent functions

As a natural extension to the rudimentary "fixed function expectation" criteria seen in §2.3.1, one can consider a family of functions $f(\cdot;\theta)$ parameterized by $\theta \in \mathbb{R}$, for which the value of θ is allowed to be determined based on the $\mathsf{L} \in \mathcal{L}$ being evaluated. While countless possibilities exist, one class of criteria that captures many important special cases in the literature is to take $f(u;\theta) = \theta + \rho(u - \theta)$, and to *optimize* with respect to θ after taking expectation. As a typical example, consider

$$\underline{\mathrm{C}}_{\rho}(\mathsf{L}) \coloneqq \inf_{\theta \in \mathbb{R}} [\theta + \mathbf{E}[\rho(\mathsf{L} - \theta)]] \tag{6}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ used in (6) is assumed to be such that for each $\mathsf{L} \in \mathcal{L}$, the function $\theta \mapsto \theta + \mathbf{E}[\rho(\mathsf{L} - \theta)]$ achieves its minimum on $\mathbb{R}$, and that $\inf\{\underline{\mathrm{C}}_{\rho}(\mathsf{L}) : \mathsf{L} \in \mathcal{L}\} > -\infty$. As we will discuss later, special cases of the learning criterion $\underline{\mathrm{C}}_{\rho}(\cdot)$, as well as close variants, arise frequently in the

literature, particularly in the case where ρ is monotonic (non-decreasing) and convex on $\mathbb{R}$. As the following result shows, under the zero-one loss, monotonicity alone is sufficient to imply collapse into error probability minimizers.

Proposition 2.5 (Collapse under monotonic dispersion)**.** *With $\underline{\mathsf{C}}_\rho$ as defined in (6), we have*

$$\mathcal{H}^*_{\mathcal{E}} \subset \arg\min_{h\in\mathcal{H}} \underline{\mathsf{C}}_\rho(\mathsf{L}_{01}(h))$$

when $\rho(\cdot)$ is non-decreasing, and equality when $\rho(\cdot)$ is increasing.

As we note in the following remarks, a wide range of well-known criteria are captured by Proposition 2.5, or can be shown to collapse using an analogous argument.

Remark 2.6 (Special case: OCE criteria)**.** When ρ in (6) is assumed to be a closed, convex, sub-differentiable function, normalized such that $\rho(0) = 0$ and $1 \in \partial\rho(0)$, which is monotonically non-decreasing on $\mathbb{R}$, the resulting learning criterion is called an *optimized certainty equivalent (OCE)* (see also Figure 4). Clearly, setting $\rho(u) = u$ recovers the expected value. Another well-known special case of OCE criteria is CVaR, which is recovered by setting $\rho(u) = \max\{0, u\}/(1-\beta)$ with $0 < \beta < 1$. Clearly this ρ is non-decreasing, so the inclusion in Proposition 2.5 holds; note that Theorem 2.2 is thus implied. On the other hand, the ρ used for CVaR is not strictly increasing. Under the zero-one loss, are there CVaR-optimal solutions that are *not* optimal in terms of error probability? This is indeed possible, but it depends on both $\mathcal{H}$ and β; we discuss this in detail in §B.2. Another well-studied special case from the OCE class is the *tilted risk* (also *entropic risk*), recovered by setting $\rho(u) = (\mathrm{e}^{\gamma u} - 1)/\gamma$ with $\gamma > 0$, taking the form

$$\underline{\mathsf{C}}_\rho(\mathsf{L}) = \frac{1}{\gamma}\log\left(\mathbf{E}\left[\mathrm{e}^{\gamma\,\mathsf{L}}\right]\right). \tag{7}$$

Note that since the exponential function is monotonically increasing, so is ρ, and thus by Proposition 2.5, tilted risk minimization under $\mathsf{L} = \mathsf{L}_{01}$ is identical to $\mathcal{E}(\cdot)$ minimization. To conclude this remark, we note that more generally, the monotonicity of all OCE risks immediately implies that all $\mathcal{E}(\cdot)$-optimal rules are also $\underline{\mathsf{C}}_\rho$-optimal, even without the other assumptions of convexity and normalization.

Remark 2.7 (Related case: Cressie-Read DRO)**.** An important class of *distributionally robust optimization (DRO)* criteria is defined in a very similar way to the OCE class given in (6). In particular, using the Cressie-Read family of f-divergences leads to criteria of the form

$$\mathrm{DRO}_{c,\varepsilon}(\mathsf{L}) := \inf_{\theta\in\mathbb{R}}\left[\theta + \left(\mathbf{E}\left[\rho_{c,\varepsilon}(\mathsf{L}_{01}(h) - \theta)\right]\right)^{1/c_*}\right]$$

where we have defined $\rho_{c,\varepsilon}(x) := (1 + c(c-1)\varepsilon)^{c_*/c}(x)_+^{c_*}$ and $c_* := c/(c-1)$, and these criteria are parameterized by $c > 1$ and $\varepsilon \geq 0$. Note that the function $\rho_{c,\varepsilon}$ is clearly non-decreasing on $\mathbb{R}$, just like in our discussion of CVaR in Remark 2.6. With $c > 1$, Cressie-Read DRO criteria are not strictly speaking OCE criteria, since the expected value is wrapped within $(\cdot)^{1/c_*}$. That said, an argument analogous to that in the proof of Proposition 2.5 using monotonicity can be used to show the same result, i.e., that $\mathcal{H}^*_{\mathcal{E}}$ is included in the solution set of $\mathrm{DRO}_{c,\varepsilon}(\mathsf{L}_{01}(\cdot))$.

Remark 2.8 (Related case: criteria based on Orlicz regret)**.** Let $f : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be a proper, lower semi-continuous convex function, satisfying $f(1) = 0$, $f(0) < \infty$, and super-coercivity in that $f(u)/u \to \infty$ as $u \to \infty$. Letting f^* denote the usual convex conjugate of f, namely $f^*(u) := \sup_{v\in\mathbb{R}}[uv - f(v)]$, in recent work by Fröhlich & Williamson (2023), a class of learning criteria are introduced with the form

$$\underline{\mathsf{C}}_{f,\varepsilon}(\mathsf{L}) := \inf_{\theta,\sigma}\left[\sigma\left(\varepsilon + \theta + \mathbf{E}\left[f^*\left(\frac{\mathsf{L}}{\sigma} - \theta\right)\right]\right)\right], \tag{8}$$

where the infimum is taken over $\theta \in \mathbb{R}$ and $\sigma > 0$, and $\varepsilon > 0$ is a parameter of the criterion. Note that f^* is finite, convex, and monotonically increasing on $\mathbb{R}$ (Fröhlich & Williamson, 2023, Prop. 3.2). Using this strong monotonicity property, just as we did for OCE criteria with increasing ρ, we can prove that under the zero-one loss, $\underline{\mathsf{C}}_{f,\varepsilon}(\mathsf{L}_{01}(\cdot))$-optimal decisions coincide with the $\mathcal{E}(\cdot)$-optimal decisions.

Remark 2.9 (Non-monotonic alternative: variantile)**.** The variance of a random variable, say L, can be naturally expressed as the minimum value of the function $\mathbf{E}(\mathsf{L} - \theta)^2$, with the minimum taken with respect to $\theta \in \mathbb{R}$. The dispersion around θ here is measured in a symmetric fashion, since the same function is used both when $\mathsf{L} > \theta$ and $\mathsf{L} \leq \theta$. A natural *asymmetric* extension considers re-scaling each of these cases with $2(1-\tau)$ and 2τ respectively, where τ is a free parameter such that $0 < \tau < 1$. Writing this explicitly as a learning criterion, we have

$$\underline{\mathsf{C}}_\tau(\mathsf{L}) := \min_{\theta\in\mathbb{R}} 2\mathbf{E}\left[|\mathbf{1}_\theta(\mathsf{L}) - \tau|(\mathsf{L} - \theta)^2\right] \tag{9}$$

where $\mathbf{1}_\theta(\mathsf{L}) := \mathrm{I}\{\mathsf{L} \leq \theta\}$. The value of θ that minimizes the function shown on the right-hand side of (9) is well-known in the economics literature as the "expectile," and while the residual $\underline{\mathsf{C}}_\tau(\mathsf{L})$ itself has received less attention, it has appeared in recent work under names such as "variancile" and "variantile" (Frongillo & Kash, 2021). Using the latter term here, the *variantile* in (9) generalizes the variance, with $\tau = 1/2$ recovering the variance as a special case. While of a similar nature to the $\underline{\mathsf{C}}_\rho$ criteria give in (6), the critical difference here is that $(\cdot)^2$ is not monotonic on $\mathbb{R}$. As a result, collapse in the sense of Proposition 3.1 and Remarks 2.6–2.8 is not guaranteed, but instead the solution set collapses into the set of *either* minimizers or maximizers of $\mathcal{E}(\cdot)$, analogous to §2.3.1. See §B.3 for a more detailed demonstration of this.

3. Relationship with Surrogate Losses

Our analysis and discussion of criterion collapse in the preceding section was centered around losses with a Bernoulli distribution, which typically arise when using the zero-one loss ℓ_{01} defined in (1). While this loss function is ubiquitous as a metric for *evaluation*, during training we rarely use ℓ_{01} to design a criterion to be optimized directly. It is far more common to introduce a surrogate function, say $\ell(h; x, y)$, that is more congenial to numerical optimization. This results in there being two distinct loss distributions for each candidate h, namely that of the binary error $\mathsf{L}_{01}(h) = \ell_{01}(h; \mathsf{X}, \mathsf{Y})$ and the surrogate loss $\mathsf{L}(h) = \ell(h; \mathsf{X}, \mathsf{Y})$. Say we introduce some new criterion map $\mathrm{C}(\cdot)$. We already know from §2 how most typical criteria collapse under $\mathsf{L}_{01}(h)$, but these insights do not apply in general for $\mathsf{L}(h)$, which may have a much more complicated distribution than a simple Bernoulli. Placing our focus on the *surrogate* criterion $\mathrm{C}(\mathsf{L}(\cdot))$, there are two natural instances of "collapse" with respect to error probability minimizers that we can conceive of:

$$\mathcal{H}^*_{\mathcal{E}} \subset \arg\min_{h \in \mathcal{H}} \mathrm{C}(\mathsf{L}(h)) \tag{10}$$

$$\arg\min_{h \in \mathcal{H}} \mathrm{C}(\mathsf{L}(h)) \subset \mathcal{H}^*_{\mathcal{E}}. \tag{11}$$

In some learning scenarios, the properties (10) and (11) may of course be desirable. In the case of (10), which is the surrogate version of the collapse notion seen in §2, we have that it is possible to be optimal in $\mathrm{C}(\mathsf{L}(\cdot))$ without having to make a sacrifice in terms of error probability $\mathcal{E}(\cdot)$. As for (11), in the traditional setup where $\mathrm{C}(\cdot) = \mathbf{E}[\cdot]$, proving that (11) holds is the central goal of "classification calibrated" surrogate design (Bartlett et al., 2006).

On the other hand, the inclusion in (11) means that error probability minimization is *unavoidable* (under $\mathrm{C}(\cdot)$ and $\ell(\cdot)$), which could have unintended repercussions in terms of other metrics such as fairness or privacy, for which tradeoffs with accuracy are well-known (Menon et al., 2021). In §3.1 we first give a simple example showing how both (10) and (11) can fail to hold under a limited model $\mathcal{H}$. In contrast, when $\mathcal{H}$ is a highly expressive model, we show in §3.2 that for most important (binary) surrogate losses and a large class of criteria (in particular the "tilted risk" criteria), average error optimizers are in fact unavoidable (i.e., (11) holds). We complement this in §3.3 by showing how criteria underlying typical "ascent-descent" gradient-based algorithms can be used to avoid unintentional collapse into $\mathcal{H}^*_{\mathcal{E}}$.

3.1. Disentangling the error and surrogate loss

For the remainder of §3, we will focus on the binary classification case, where our random data takes the form $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$, taking values in $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$. Let $\mathcal{S}$ denote a set of scoring functions $s : \mathcal{X} \to \mathbb{R}$, with decision set $\mathcal{H} := \{\mathrm{sign}(s(\cdot)) : s \in \mathcal{S}\}$ based on $\mathcal{S}$, noting that for any $u \in \mathbb{R}$, $\mathrm{sign}(u)$ is 1 when $u \geq 0$, and is -1 when $u < 0$. In this setting, we show how under a model $\mathcal{H}$ with limited expressive power, it is very easy to construct an example where both notions of "collapse" just mentioned cannot hold.

For simplicity and ease of visualization, consider an example on the plane, i.e., where $\mathcal{X} = \mathbb{R}^2$. Let $\mathsf{X} = (\mathsf{V}_1, \mathsf{V}_2)$, where V_1 and V_2 are real-valued random variables. Fixing an arbitrary $a > 1$, let V_1 take values in $\{-1, 1, a\}$ and let V_2 take values in $\{-a, -1, 1\}$. Assume that $\mathbf{P}\{\mathsf{X} = (-1, 1)\} = \mathbf{P}\{\mathsf{X} = (1, -1)\} > 0$, and writing

$$p := \mathbf{P}\{\mathsf{X} = (-1, 1)\} + \mathbf{P}\{\mathsf{X} = (1, -1)\}$$

let us say that $\mathbf{P}\{\mathsf{X} = (a, -a)\} = 1 - p$ and $p < 1$ also hold. As for the labeling, let us assume

$$\mathsf{Y} = \mathrm{sign}(\sqrt{2} - \|\mathsf{X}\|_2)\,\mathrm{sign}(\mathsf{V}_2 - \mathsf{V}_1). \tag{12}$$

Set $\mathcal{S} := \{s_1, s_2\}$, where $s_1(v_1, v_2) := v_2 - v_1$ and $s_2(v_1, v_2) := v_1 - v_2$ for all $v_1, v_2 \in \mathbb{R}$. The resulting classifiers are $h_1(x) := \mathrm{sign}(s_1(x))$ and $h_2(x) := \mathrm{sign}(s_2(x))$ for $x \in \mathbb{R}^2$. See Figure 1 for a visualization of the three possible data points and the role of p. Error probabilities are $\mathbf{P}\{h_1(\mathsf{X}) \neq \mathsf{Y}\} = 1 - p$ and $\mathbf{P}\{h_2(\mathsf{X}) \neq \mathsf{Y}\} = p$. When $p > 1/2$, clearly $\mathcal{H}^*_{\mathcal{E}} = \{h_1\}$. As a typical surrogate loss, let us consider the usual binary "logistic loss," namely the binary cross-entropy loss with scores converted to probabilities via the logistic function. Losses take the form $\ell_{\log}(s; x, y) := \log(1 + \exp(-ys(x)))$ for each $x \in \mathbb{R}^2$ and $y \in \{-1, 1\}$. Writing $\mathsf{L}_{\log}(s) := \ell_{\log}(s; \mathsf{X}, \mathsf{Y})$, the surrogate loss distributions that result are quite simple; some arithmetic shows that we have

$$\mathsf{L}_{\log}(s_1) \in \{\log(1 + \exp(-2)), \log(1 + \exp(2a))\},$$
$$\mathsf{L}_{\log}(s_2) \in \{\log(1 + \exp(-2a)), \log(1 + \exp(2))\}$$

with probability 1. For any fixed value of $a > 1$, it is always possible to take $0 < p < 1$ close enough to 1 that in terms of the *expected* surrogate loss, we have

$$\arg\min_{s \in \mathcal{S}} \mathbf{E}[\mathsf{L}_{\log}(s)] = \{s_1\}$$

which is in agreement with the error probability minimizer. On the other hand, regardless of what value of $0 < p < 1$ is assumed, both the *maximum* and *minimum* losses incurred by s_2 are respectively smaller than those of s_1. That is,

$$\max_{\mathsf{X},\mathsf{Y}} \mathsf{L}_{\log}(s_2) < \max_{\mathsf{X},\mathsf{Y}} \mathsf{L}_{\log}(s_1),$$
$$\min_{\mathsf{X},\mathsf{Y}} \mathsf{L}_{\log}(s_2) < \min_{\mathsf{X},\mathsf{Y}} \mathsf{L}_{\log}(s_1).$$

This simple observation means that criteria which tend to emphasize optimizing either the "worst case" or "best case"

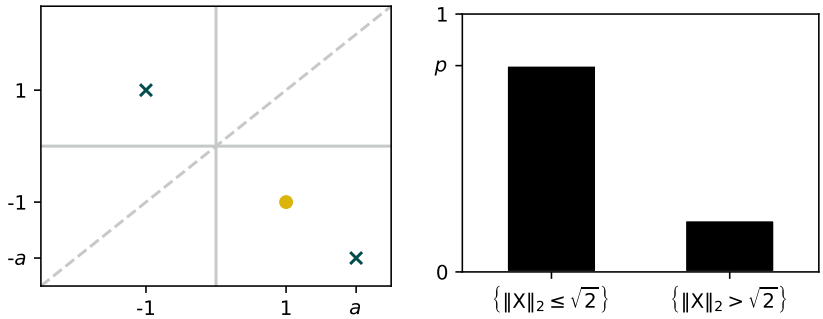

Figure 1. In the left plot, we show the three possible data points that can arise in the example described in §3.1. Points above the dashed silver line are assigned a label of 1 by h_1 and -1 by h_2; signs are reversed for all points below this line. For the outlying point in the bottom right, we have set $a = 2$ in this example. In the right plot, we illustrate setting $p > 1/2$ to ensure the optimality of h_1 and h_2 in distinct criteria diverges.

in terms of surrogate loss values will end up disagreeing with the error probability $\mathcal{E}(\cdot)$, and thus neither (10) nor (11) can hold. This is not limited to the extreme case where the criterion map is $\mathrm{C}(\cdot) = \max_{\mathsf{X},\mathsf{Y}}(\cdot)$ or $\mathrm{C}(\cdot) = \min_{\mathsf{X},\mathsf{Y}}(\cdot)$. For example, analogous results can easily be shown to hold under the γ-tilted risk $\mathsf{L} \mapsto (1/\gamma)\log(\mathbf{E}[\mathrm{e}^{\gamma\,\mathsf{L}}])$ (recalling Remark 2.6) for $\gamma \neq 0$ with $|\gamma|$ sufficiently large, with $\gamma < 0$ emphasizing the best case, and $\gamma > 0$ emphasizing the worst case.

3.2. Criteria that cannot avoid collapse

Let us assume that ℓ is a surrogate loss function which applies penalties to classification margins, i.e., assume that $\ell(s; x, y) := \phi(ys(x))$, where margin penalizer $\phi : \mathbb{R} \to \mathbb{R}_+$ is assumed to be convex and non-negative over $\mathbb{R}$, as well as differentiable at 0; one typical example is $\ell = \ell_{\log}$ from §3.1. Let us also generalize beyond the expected loss criteria by considering the optimized certainty equivalent (OCE) criterion, defined for any $\mathsf{L} \in \mathcal{L}$ by

$$\mathrm{OCE}(\mathsf{L}) := \inf_{\theta \in \mathbb{R}} \left[\theta + \mathbf{E}[\rho(\mathsf{L} - \theta)]\right] \tag{13}$$

and characterized by $\rho : \mathbb{R} \to \mathbb{R}$ satisfying the properties mentioned in Remark 2.6. Our running assumption is that for each $\mathsf{L} \in \mathcal{L}$, there is always a $\theta^* \in \mathbb{R}$ such that $\mathrm{OCE}(\mathsf{L}) = \theta^* + \mathbf{E}[\rho(\mathsf{L} - \theta^*)]$ holds. We show below that for a large sub-class of OCE criteria and loss functions, criteria collapse in the sense of (11) often cannot be avoided.

Proposition 3.1. *Assume that the margin penalizer satisfies $\phi'(0) < 0$, and that ρ in (13) is increasing (not just non-decreasing) and differentiable. With $\mathcal{H} := \{\mathrm{sign}(s(\cdot)) : s \in \mathcal{S}\}$, let $\mathcal{S}$ be the set of all measurable functions on $\mathcal{X}$, and let $(s_1, s_2, \ldots)$ be a sequence from $\mathcal{S}$, with the resulting classifiers denoted by $h_n(\cdot) := \mathrm{sign}(s_n(\cdot))$. Write $\mathsf{L}_\phi(h) := \ell(s_h; \mathsf{X}, \mathsf{Y}) = \phi(\mathsf{Y} s_h(\mathsf{X}))$ for the losses induced by ϕ. Taking $n \to \infty$, the following implication holds:*

$$\mathrm{OCE}(\mathsf{L}_\phi(h_n)) \to \inf_{h \in \mathcal{H}} \mathrm{OCE}(\mathsf{L}_\phi(h)) \implies \mathcal{E}(h_n) \to \mathcal{E}^*,$$

where $\mathcal{E}^* := \inf_{h \in \mathcal{H}} \mathcal{E}(h)$. In addition, minimizers of the OCE surrogate satisfy

$$\arg\min_{h \in \mathcal{H}} \mathrm{OCE}(\mathsf{L}_\phi(h)) \subset \mathcal{H}^*_{\mathcal{E}}.$$

Remark 3.2 (Classification calibrated surrogates)**.** Our assumptions on the margin penalizer $\phi(\cdot)$ in Proposition 3.1 amount to it being *classification calibrated* in the sense of Bartlett et al. (2006), a basic property common in all popular surrogates for binary classification. Many choices of $\phi(\cdot)$ are allowed by Proposition 3.1: in addition to monotonic losses such as the logistic loss $\phi(u) = \log(1 + \exp(-u))$, exponential loss $\phi(u) = \exp(-u)$, and hinge loss $\phi(u) = \max\{0, 1-u\}$, it also captures non-monotonic choices such as the quadratic loss $\phi(u) = (1-u)^2$, the ARC-X4 loss $\phi(u) = |1-u|^5$ (Breiman, 1999), and shifted Huber-Catoni penalties (Holland, 2019).

Remark 3.3 (Strict monotonicity of ρ)**.** While the class of OCE criteria in its most general form just asks for ρ to be non-decreasing, in Proposition 3.1 we require that it be strictly monotonic (increasing). Should a flat (zero slope) region be allowed to exist, it allows for $\mathrm{OCE}(\mathsf{L}_{\log}(\cdot))$-optimal but $\mathcal{E}(\cdot)$-sub-optimal solutions to exist. This is akin to the "positive steepness" condition seen in related work on DRO (Hu et al., 2018, Lem. 1). This means that the special case of CVaR, amounting to $\mathrm{OCE}(\cdot)$ with $\rho(u) = \max\{0, u\}/(1-\beta)$ and $0 < \beta < 1$, is excluded. On the other hand, the closely related γ-tilted risk, with $\rho(u) = (\mathrm{e}^{\gamma u} - 1)/\gamma$ and $\gamma > 0$, is included. Under models with high expressive capacity, Proposition 3.1 suggests that methods such as tilted ERM may not always be a trustworthy choice for learning tasks where some evaluation metric of interest (e.g., fairness, interpretability) is at odds with error probability.

3.3. Criteria that can avoid collapse

Here we consider alternatives to the class of criteria captured by the "unavoidability" result in Proposition 3.1. To do so, we introduce minor changes to the functional form of the OCE criterion of (13), and highlight links to existing algorithms in the literature (see Remark 3.5). Let $\widetilde{\rho} : \mathbb{R} \to [0, \infty)$ be a continuous convex function which is coercive (i.e., $|u| \to \infty$ implies $\widetilde{\rho}(u) \to \infty$), takes its minimum at 0 (assuming $\widetilde{\rho}(0) = 0$), and is strictly convex on a neighborhood of 0. We use the symbol $\widetilde{\rho}$ (instead of ρ as in §3.2) because these assumptions imply that $\widetilde{\rho}$ cannot be monotonic on $\mathbb{R}$. Let us denote the *inner* and *outer loss-restraining criterion maps* by $\mathrm{C}_{\text{inn}}(\cdot)$ and $\mathrm{C}_{\text{out}}(\cdot)$ respectively, defined for any $\mathsf{L} \in \mathcal{L}$ by

$$\mathrm{C}_{\text{inn}}(\mathsf{L}; \theta) \coloneqq \widetilde{\rho}(\mathbf{E}[\mathsf{L}] - \theta), \tag{14}$$

$$\mathrm{C}_{\text{out}}(\mathsf{L}; \theta) \coloneqq \mathbf{E}[\widetilde{\rho}(\mathsf{L} - \theta)]. \tag{15}$$

See Figure 4 for some examples of valid $\widetilde{\rho}$ compared with valid ρ used for OCE criteria. Intuitively, $\mathrm{C}_{\text{inn}}(\mathsf{L}; \theta)$ simply asks the learning algorithm to arrive at a loss distribution whose mean is "close" to some threshold θ; when $\theta = 0$, $\mathrm{C}_{\text{inn}}(\mathsf{L}; \theta)$ and $\mathbf{E}[\mathsf{L}]$ are equivalent (i.e., their solution sets coincide). In contrast, $\mathrm{C}_{\text{out}}(\mathsf{L}; \theta)$ asks that the loss distribution be *well-concentrated* near θ. In both cases, the notions of "closeness" and "concentration" are broad, and can be asymmetric (e.g., heavier right tails than left tails are allowed) when $\widetilde{\rho}$ is. The following result shows how even under a highly expressive model as in §3.2, constraining the loss distribution by means of criteria such as C_{inn} and C_{out} lets us circumvent the unavoidability of Proposition 3.1.

Proposition 3.4. *Under the assumptions of Proposition 3.1, assume that $\mathcal{E}^* = 0$ and the margin penalizer $\phi(\cdot)$ is monotonic (non-increasing) and unbounded above. Consider the non-trivial case where $\mathcal{H}^*_{\mathcal{E}} \neq \mathcal{H}$. Then, there exists $\theta > 0$ such that (11) cannot hold under either of the loss-restraining criteria, i.e., we have*

$$\mathcal{H}^*_{\mathcal{E}} \cap \arg\min_{h \in \mathcal{H}} \mathrm{C}(\mathsf{L}_\phi(h); \theta) = \emptyset$$

for both $\mathrm{C} = \mathrm{C}_{\text{inn}}$ and $\mathrm{C} = \mathrm{C}_{\text{out}}$.

Remark 3.5 (Links to algorithms in the literature)**.** While we have introduced the loss-restraining criteria C_{inn} and C_{out} in (14) and (15) in a rather general form, plugging in concrete examples of $\widetilde{\rho}$ can be shown to align with specific learning algorithms from the literature. Perhaps the best-known is that of "Flooding," as studied by Ishida et al. (2020). By setting $\widetilde{\rho}(\cdot) = |\cdot|$ and applying sub-gradient descent with step-size $\alpha \geq 0$ to the resulting $\mathrm{C}_{\text{inn}}(\mathsf{L}(\cdot); \theta)$ for any differentiable loss L, we end up with an update rule

$$h_{t+1} = h_t - \alpha \operatorname{sign}(\mathbf{E}[\mathsf{L}(h_t)] - \theta)\mathbf{E}[\nabla \mathsf{L}(h_t)]. \tag{16}$$

In practice, of course, the true expectation is replaced with the empirical mean over a finite sample. From (16) it is clear that as the sequence $(h_1, h_2, \ldots)$ progresses, as long as the expected loss is above threshold θ, vanilla gradient descent (on the expected loss) is run. It is only if the expected loss falls below this threshold that the sign flips, changing the update to gradient *ascent*. The original intuition behind this approach was simply to avoid over-optimizing a surrogate loss function, but since the original paper, links between this method and sharpness-aware minimization (Foret et al., 2021) have been established; see for example Karakida et al. (2023). The hard condition for switching between ascent and descent was softened in recent work by Holland & Nakatani (2023); their approach can be interpreted in our framework as using C_{out} rather than C_{inn}, and replacing the absolute value function with $\widetilde{\rho}(\cdot) = \sqrt{(\cdot)^2 + 1} - 1$. The resulting gradient descent update looks like

$$h_{t+1} = h_t - \alpha \mathbf{E}[\widetilde{\rho}'(\mathsf{L}(h_t) - \theta)\nabla \mathsf{L}(h_t)]. \tag{17}$$

This procedure is called "softened ascent-descent" (SoftAD) by the authors; note that the hard switch of $\operatorname{sign}(\cdot)$ in (16) is replaced in (17) by a smooth switch $\widetilde{\rho}'(\cdot)$, plus the modulation is applied in a per-point fashion before averaging, rather than after. Given that the Flooding and SoftAD methods just described are captured as special cases in Proposition 3.4, we shall pay particular attention to them in our empirical analysis to follow in §4.

Remark 3.6 (Alternative approaches in the literature)**.** In the context of avoiding collapse of DRO criteria (e.g., Theorem 2.1), Hu et al. (2018, §4) consider constraining the uncertainty set (our $\mathcal{P}$) such that the underlying data distribution is effectively controlled by a simple discrete latent variable. This nice trick allows them to express their "structural ARM" criterion in terms of a weighted sum of expected zero-one loss and variance-like terms (e.g., their equation 18), which in principle makes it possible to avoid collapse. Another clever approach to avoiding such collapse (this time for CVaR; Theorem 2.2) is to consider the task of learning an ensemble of k candidates as done by Zhai et al. (2021, §3), i.e., choosing $\mathbf{h} \coloneqq (h_1, \ldots, h_k)$ with each $h_j \in \mathcal{H}$ and a weight vector $\lambda \coloneqq (\lambda_1, \ldots, \lambda_k)$. The underlying *loss function* is then a weighted sum of zero-one losses, namely $\ell_{01}^{(k)}(\mathbf{h}, \lambda; x, y) \coloneqq \sum_{j=1}^{k} \lambda_j \ell_{01}(h_j; x, y)$, and thus in principle we can avoid collapse since the distribution is no longer Bernoulli. While both of these results describe settings in which collapse *need not* occur under $\mathsf{L} = \mathsf{L}_{01}$, collapse is still possible, and there are not conditions for avoiding collapse with surrogates (as in our Proposition 3.4).

4. Empirical Analysis

Within our theoretical analysis of §3, we contrasted learning criteria that can and cannot avoid collapse into error probability minimizers. In particular, we emphasized the role of monotonicity in measuring the dispersion of losses around

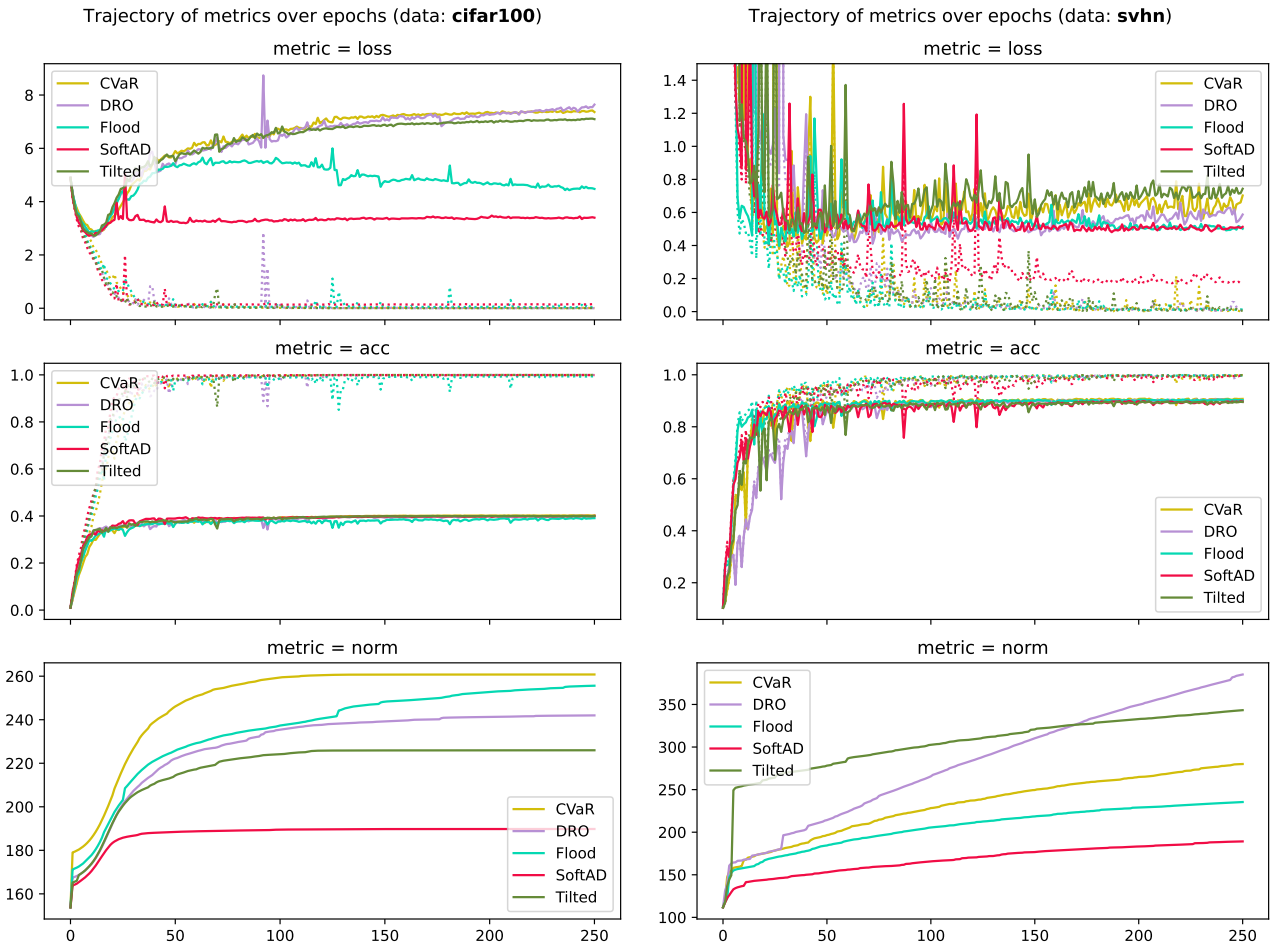

Figure 2. Key metrics of interest (vertical axis) over epochs (horizontal axis). Here "`loss`" refers to average surrogate loss, "`acc`" refers to accuracy, and "`norm`" refers to the model L2 norm. Loss and accuracy are given for both training (dotted lines) and test data (solid lines). Plots on the left are for CIFAR-100, and plots on the right are for SVHN.

	CIFAR-10	CIFAR-100	FashionMNIST	SVHN
CVaR	0.14 (0.15)	0.28 (0.13)	0.46 (0.13)	0.20 (0.12)
DRO	0.02 (0.04)	0.0 (0.0)	0.0 (0.0)	0.17 (0.20)
Flood	0.05 (0.06)	0.03 (0.05)	0.01 (0.0)	0.01 (0.0)
SoftAD	0.06 (0.05)	0.15 (0.13)	0.01 (0.0)	0.19 (0.19)
Tilted	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)

Table 1. Hyperparameter values selected for each method to maximize validation accuracy (averaged over trials). Standard deviation (again over trials) is given in parentheses.

a threshold (ρ vs. $\widetilde{\rho}$). DRO, CVaR, and tilted ERM are typical monotonic methods, whereas Flooding and SoftAD are representative non-monotonic methods. To complement the points established in the previous section, here we implement the aforementioned methods and apply them to the training of non-linear neural network models for classification. Our interest is in *tradeoffs*. Namely, having optimized for their respective learning criterion (under a base surrogate loss), we compare performance in terms of three alternative metrics, namely average surrogate loss, accuracy, and model complexity measured by the norm of model parameters.

Setup Our experiment design essentially follows the experimental design described by Ishida et al. (2020) (who

first proposed the Flooding technique). The basic learning task is image classification from scratch. We use four standard datasets: CIFAR-10, CIFAR-100, FashionMNIST, and SVHN, accessed via the `torchvision` package. Each of these datasets has a default training-test split, which we leave fixed across all randomized trials. In each trial, however, we shuffle the training data and do an 80-20 split for training-validation. For CIFAR-10/100, we use ResNet-34; for FashionMNIST we flatten the images and use a simple feed-forward network (one hidden layer, 1000 units, batch normalization, ReLU activation); finally, for SVHN, we use ResNet-18. As an optimizer, we use vanilla SGD with step size 0.1 and momentum 0.9, run for 250 epochs, with mini-batch size of 200. In all cases, no pre-training is done; initial weights are randomly determined anew for each trial, and we note that the initial weights for any given trial are common across all methods being tested. The "base" surrogate loss function used for training is the multi-class cross-entropy loss.

Code A GitHub repository with code and seed values to re-create all the results presented in this paper is available at this URL: `https://github.com/feedbackward/collapse`.

Methods We are comparing the impact of different learning criteria, so the choice of "method" here amounts to choosing from the following: CVaR, DRO, Flooding, SoftAD, and tilted ERM. For DRO, we use Cressie-Read divergences (Remark 2.7), with $c = 2$ fixed and ε as a hyperparameter. For CVaR (cf. Theorem 2.2), the quantile level β is the hyperparameter. For both CVaR and DRO, we optimize the threshold θ simultaneously with model parameters. Tilted ERM has a closed form (7) that we use, with hyperparameter γ. For the non-monotonic methods (Remark 3.5), Flooding is $\mathrm{C}_{\mathrm{inn}}$ with $\widetilde{\rho}(u) = |u|$, and SoftAD is $\mathrm{C}_{\mathrm{out}}$ with $\widetilde{\rho}(u) = \sqrt{u^2+1} - 1$; both have threshold θ as a hyperparameter. For each method, we consider ten distinct hyperparameter candidates, as follows. CVaR: β ranges between 0 and 0.9. DRO: with $\varepsilon = (1/(1-\widetilde{\varepsilon}) - 1)^2/2$, $\widetilde{\varepsilon}$ ranges between 0 and 0.9. Flooding: θ ranges between 0.01 and 1.0. SoftAD: θ ranges between 0.01 and 0.75. Tilted ERM: γ ranges between 0 and 2.0. We remark that vanilla ERM is captured as a special case of the three monotonic methods ($\beta = 0$, $\varepsilon = 0$, and $\gamma = 0$).

Results For evaluation, before and after each epoch, we record three metrics: average surrogate loss, average accuracy (one minus average zero-one loss), and the L2 norm of model parameters. Here "average" refers to taking an average over each relevant data set (training, validation, test). Furthermore, we have run five independent trials, and the results to be presented are averaged over all trials. To choose a representative hyperparameter setting for each method, we select the value which maximizes the *validation accuracy*. In Figures 2 and 5 we show the trajectory of our metrics of interest over epochs (with epoch 0 being the initial state). In Table 1, we show the hyperparameter values that were selected for each method (here also, averaged over trials).

Discussion There are several salient points that can be extracted from the results just described. First, it is clear that in terms of accuracy, all methods that we have tested include hyperparameter settings which can achieve near-perfect training accuracy, with test accuracy essentially the same across all methods. Note that for all trials and all datasets, tilted ERM is such that optimizing for validation accuracy leads to a setting of $\gamma = 0$, namely vanilla ERM. Second, while the "best accuracy" levels are essentially uniform across the methods tested, from the perspective of *tradeoffs*, we see that things are far less uniform. In terms of average surrogate loss and model norm metrics, there is substantial dispersion between methods. We see that the two non-monotonic methods tend to achieve a far better average *test* surrogate loss than the monotonic methods, with SoftAD being unique in terms of realizing a small model norm as well. Considering these trends (in Figures 2 and 5) together with the selection trends (in Table 1), the results seem to suggest that even weakly constraining the surrogate loss distribution via $\mathrm{C}_{\mathrm{inn}}$ and $\mathrm{C}_{\mathrm{out}}$ (corresponding to Flooding and SoftAD here) looks to offer an appealing balance between the metrics we have studied here. Whether or not this behavior can be extended in a straightforward manner to metrics from other domains (e.g., fairness, prediction calibration, etc.) remains to be seen.

5. Concluding Remarks

In this work, we have formulated and studied the notion of criterion collapse in machine learning, with particular emphasis placed on settings in which *unintended* collapse may occur, and how this phenomenon changes when surrogate functions are introduced into the picture. By highlighting situations in which unintentional collapse is possible, we are hoping to slowly but surely build a methodological roadmap of sorts, highlighting potential risks with model/algorithm design decisions, something analogous to the "AI model cards" that have attracted interest in recent years (Liang et al., 2024). By complementing our theory with experiments, we try to highlight how learning algorithms derived from distinct criteria classes (here the key difference is monotonicity) can lead to rather different algorithm behavior and learned model properties, both during and after training. The idea is to take these empirical insights together with the aforementioned basic principles to guide smart AI system design when our goals are more diverse than just "maximize the accuracy on a big training set."

Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR21C6, and a grant from the SECOM Science and Technology Foundation.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. While there are many potential societal consequences of forwarding this field, these are already well-established and are not specific to this work. As such, there are no particular concerns we feel must be highlighted here.

References

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

Bertsekas, D. P. *Convex Optimization Algorithms*. Athena Scientific, 2015.

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

Curi, S., Levy, K. Y., Jegelka, S., and Krause, A. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 1036–1047, 2020.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Fröhlich, C. and Williamson, R. C. Tailoring to the tails: Risk measures for fine-grained tail sensitivity. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=UntUoeLwwu.

Frongillo, R. M. and Kash, I. A. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, 2021.

Holland, M. J. Classification using margin pursuit. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 712–720, 2019.

Holland, M. J. Learning with risks based on M-location. *Machine Learning*, 111:4679–4718, 2022. doi: 10.1007/s10994-022-06217-5.

Holland, M. J. and Nakatani, K. Implicit regularization via soft ascent-descent. *arXiv preprint arXiv:2310.10006v1*, 2023.

Holland, M. J. and Tanabe, K. A survey of learning criteria going beyond the usual risk. *Journal of Artificial Intelligence Research*, 73:781–821, 2023.

Hu, S., Wang, X., and Lyu, S. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 2023.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2029–2037, 2018.

Ishida, T., Yamane, I., Sakai, T., Niu, G., and Sugiyama, M. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4604–4614, 2020.

Karakida, R., Takase, T., Hayase, T., and Osawa, K. Understanding gradient regularization in deep learning: Efficient finite-difference computation and implicit bias. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15809–15827, 2023.

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

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

Li, T., Beirami, A., Sanjabi, M., and Smith, V. Tilted empirical risk minimization. In *The 9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=K5YasWXZT3O.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24:1–79, 2023.

Liang, W., Rajani, N., Yang, X., Ozoani, E., Wu, E., Chen, Y., Smith, D. S., and Zou, J. What's documented in AI? systematic analysis of 32K AI model cards. *arXiv preprint arXiv:2402.05160*, 2024.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

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

Reid, M. D. and Williamson, R. C. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Reid, M. D. and Williamson, R. C. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

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

Rockafellar, R. T. and Uryasev, S. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013.

Royset, J. O. Risk-adaptive approaches to learning and decision making: A survey. *arXiv preprint arXiv:2212.00856*, 2022.

Rudin, W. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

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

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

Williamson, R. C., Vernet, E., and Reid, M. D. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

Zhai, R., Dan, C., Suggala, A., Kolter, Z., and Ravikumar, P. Boosted CVaR classification. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 21860–21871, 2021.

A. Additional Notes

Tradeoffs between metrics The navigation between "measures" and their "proxies" is a well-known challenge in the field of public policy design, and has recently been highlighted as an important challenge by researchers at OpenAI.[1]

Surveys on learning criteria In recent years, several in-depth surveys looking at trends in criterion design for machine learning and related disciplines have appeared. See Holland & Tanabe (2023), Hu et al. (2023), and Royset (2022) for representative work.

OCE and DRO references See Lee et al. (2020) and Holland & Tanabe (2023, §3.2) for background on OCE criteria in machine learning, and Ben-Tal & Teboulle (2007) for an authoritative initial reference. Tilted ERM with $\gamma > 0$ is a special case of OCE, but it should be noted that a more general notion of "tilted ERM" can be considered by using the right-hand side of (7) with general $\gamma \neq 0$ (Li et al., 2021). For $\gamma < 0$ the link to OCE criteria breaks down (reducing the role of right-side tails instead of emphasizing them), but leads to an array of different applications (Li et al., 2023). For the well-known special case of CVaR, see Rockafellar & Uryasev (2000) for the landmark paper that presents the form of CVaR used in our Theorem 2.2. See Kashima (2007) and Takeda & Sugiyama (2008) for early links to machine learning. CVaR can be derived as a limiting case of DRO under a specific f-divergence ((Duchi & Namkoong, 2021, Ex. 3)). For more general results on DRO with uncertainty sets constructed using on f-divergences, see Namkoong & Duchi (2016) and Duchi & Namkoong (2021).

Tail-sensitive criteria The basic argument of §3.1 can be extended well beyond the tilted ERM criterion to a wide variety of other criteria including many OCE risks and their "inverted" variants; see Lee et al. (2020) or Holland & Tanabe (2023) for some representative examples of such criteria.

References on surrogate design There is a very large literature on the topic of designing surrogate functions under the $\mathbf{E}[\cdot]$ criterion. For binary classification, see for example Bartlett et al. (2006) and Reid & Williamson (2010; 2011). For the multi-class setting, see Williamson et al. (2016), and the references within.

Remark A.1 (Coherent risks)**.** One of the most well-known and well-studied classes of risk functions is that of "coherent risk" criteria.[2] Coherent risks are characterized by several properties, including a very weak notion of monotonicity, which simply requires that for any random losses L and L', a criterion map $\mathrm{C}(\cdot)$ satisfies the following:

$$\mathsf{L} \leq \mathsf{L}' \text{ (almost surely)} \implies \mathrm{C}(\mathsf{L}) \leq \mathrm{C}(\mathsf{L}'). \tag{18}$$

While a very natural property in a general context, when we restrict ourselves to Bernoulli losses, with say $\mathsf{L} = \mathsf{L}_{01}(h_1)$ and $\mathsf{L}' = \mathsf{L}_{01}(h_2)$ based on candidates $h_1, h_2 \in \mathcal{H}$, this monotonicity property becomes rather vacuous. Just consider when the condition in (18) actually can hold. It can only be valid in two cases: either we have $\mathsf{L} = \mathsf{L}'$ (almost surely) or $\mathsf{L} < \mathsf{L}'$ (almost surely). Note that for $\mathsf{L} < \mathsf{L}'$ to hold almost surely, we must have $\mathsf{L} = 0$ and $\mathsf{L}' = 1$ almost surely. In any case, we are dealing with either constant random variables or identical distributions, so the notion of "monotonicity" only really becomes interesting for more complicated distributions.

Gaps between theory and practice? As discussed in §3.2, our Proposition 3.1 could be interpreted as a "negative result" showing that criteria like tilted ERM (with positive tilt) cannot avoid error probability minimization. On the other hand, such criteria have been applied with effect to classification under class imbalance learning problems (e.g., Li et al. (2023, §7.2.4)). We remark that there is no conflict between our results and such empirical results, since in practice, we only have limited expressive power and more importantly limited capability to optimize a typically non-convex objective (viewed as a function on $\mathcal{H}$). When effective model capacity is constrained, it is easy to find settings in which the zero-one loss mean is not controlled by relevant surrogate loss criterion (e.g., our example from §3.1).

Notes on experiment results Looking at Table 1, note that for CIFAR-100 and FahsionMNIST, in all five trials, the hyperparameter setting for both DRO and Tilted ERM was zero, i.e., vanilla ERM was selected. In the case of FashionMNIST, the trajectories in Figure 5 (right plot) align perfectly, but for CIFAR-100 (Figure 2, left plot), they do not; this is due to randomness inherent in the PyTorch implementation of ResNet-34 (no pre-training) that we have left uncontrolled.

[1] https://openai.com/research/measuring-goodharts-law

[2] See Rockafellar & Uryasev (2013, §3) for background.

B. Proofs

B.1. Proofs of results given in the main text

Proof of Proposition 2.3. Fixing $h \in \mathcal{H}$ and taking $0 < \beta \leq 1 - \mathcal{E}(h)$, we know that for any choice of $x \geq 0$ we have $\mathbf{P}\{\mathsf{L}_{01}(h) \leq x\} \geq \beta$, but this fails for any $x < 0$. Thus we have $\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 0$ for any β in this range. For larger probability levels, namely where $1 - \mathcal{E}(h) < \beta \leq 1$, we know that $\mathbf{P}\{\mathsf{L}_{01}(h) \leq x\} \geq \beta$ holds for any $x \geq 1$, but fails for any $x < 1$, giving us $\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 1$ for this range. Taking these facts together, we have

$$\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = \begin{cases} 1, & \text{if } 1 - \mathcal{E}(h) < \beta \leq 1 \\ 0, & \text{if } 0 < \beta \leq 1 - \mathcal{E}(h). \end{cases}$$

Now, consider the task of minimizing $\mathrm{Q}_\beta(\mathsf{L}_{01}(h))$ as a function of $h \in \mathcal{H}$ given a fixed level $0 < \beta < 1$. Note that in all situations, we can always say that

$$\mathcal{H}^*_{\mathcal{E}} \subset \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathrm{Q}_\beta(\mathsf{L}_{01}(h)). \tag{19}$$

This should be intuitively clear since achieving a small quantile amounts to achieving a sufficiently small error probability, though the probability need not be *minimal*. For completeness, let us deal with the corner cases. If $\min_{h \in \mathcal{H}} \mathcal{E}(h) > 1 - \beta$, then regardless of what h we choose, the value is $\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 1$, and so every $h \in \mathcal{H}$ is trivially "optimal." The same trivial optimality arises if $\max_{h \in \mathcal{H}} \mathcal{E}(h) \leq 1 - \beta$, since all h result in $\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 0$. When neither of these conditions hold, choosing $h \in \mathcal{H}$ such that the error probability is small enough gives us the desired minimum quantile value. In all cases, the inclusion (19) holds. □

Proof of Proposition 2.5. For convenience, let us write

$$\mathrm{C}_\rho(\mathsf{L}; \theta) \coloneqq \theta + \mathbf{E}[\rho(\mathsf{L} - \theta)]$$

for any $\mathsf{L} \in \mathcal{L}$ and $\theta \in \mathbb{R}$. Since our running assumption is that the minimum (in θ) is achieved on $\mathbb{R}$, we have that $\underline{\mathrm{C}}_\rho(\mathsf{L}) = \min_{\theta \in \mathbb{R}} \mathrm{C}_\rho(\mathsf{L}; \theta)$. Let $h_1, h_2 \in \mathcal{H}$ be any candidates such that $\mathcal{E}(h_1) \leq \mathcal{E}(h_2)$. In addition, let θ_1 and θ_2 respectively minimize $\mathrm{C}_\rho(\mathsf{L}_{01}(h_1); \cdot)$ and $\mathrm{C}_\rho(\mathsf{L}_{01}(h_2); \cdot)$. With these assumptions in place, the following chain of inequalities holds:

$$\begin{aligned}
\underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_1)) &= \mathrm{C}_\rho(\mathsf{L}_{01}(h_1); \theta_1) \\
&\leq \mathrm{C}_\rho(\mathsf{L}_{01}(h_1); \theta_2) \\
&= \theta_2 + \rho(-\theta_2) + \mathcal{E}(h_1)\left(\rho(1 - \theta_2) - \rho(-\theta_2)\right) \\
&\leq \theta_2 + \rho(-\theta_2) + \mathcal{E}(h_2)\left(\rho(1 - \theta_2) - \rho(-\theta_2)\right) \\
&= \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_2)).
\end{aligned}$$

Note that the second inequality uses the assumed monotonicity of ρ, which implies $\rho(1 - \theta_2) \geq \rho(-\theta_2)$. Since the choice of h_1 and h_2 was arbitrary, we can immediately conclude for any $h_1, h_2 \in \mathcal{H}$ that

$$\mathcal{E}(h_1) \leq \mathcal{E}(h_2) \implies \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_1)) \leq \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_2)).$$

Another immediate conclusion is that

$$\mathcal{H}^*_{\mathcal{E}} \subset \operatorname*{arg\,min}_{h \in \mathcal{H}} \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h)). \tag{20}$$

If ρ is monotonically *increasing* (rather than just non-decreasing), then it follows that

$$\mathcal{E}(h_1) < \mathcal{E}(h_2) \implies \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_1)) < \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h_2)).$$

This means that one cannot be sub-optimal in $\mathcal{E}(\cdot)$ while still being optimal in $\underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(\cdot))$. As such, the two solution sets must coincide:

$$\mathcal{H}^*_{\mathcal{E}} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h)). \tag{21}$$

Taking the conditions for (20) and (21), we have the desired result. □

Proof of Proposition 3.1. To get started, we would like to express the OCE criterion in terms of a modified penalty applied to binary classification margins. Let us denote the modified penalty based on ϕ by

$$\widetilde{\phi}(u;\theta) \coloneqq \theta + \rho(\phi(u) - \theta) \tag{22}$$

with the corresponding θ-dependent expectation denoted by

$$\mathrm{C}_\phi(s;\theta) \coloneqq \mathbf{E}[\widetilde{\phi}(\mathsf{Y}s(\mathsf{X});\theta)]. \tag{23}$$

By our running assumptions, there always exists a minimizer of $\theta \mapsto \mathrm{C}_\phi(s;\theta)$ on $\mathbb{R}$, which for convenience we will denote by

$$\theta^*(s) \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}} \mathrm{C}_\phi(s;\theta). \tag{24}$$

With this notation in hand, it follows that for each $s \in \mathcal{S}$ we have

$$\mathrm{C}_\phi(s;\theta^*(s)) = \mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))). \tag{25}$$

Moving forward, we will want to study sequences $(s_1, s_2, \ldots)$ of functions in $\mathcal{S}$ that minimize the OCE criterion in terms of the surrogate loss. More formally, let us assume $(s_1, s_2, \ldots)$ is any sequence satisfying

$$\lim_{n\to\infty} \mathrm{OCE}(\phi(\mathsf{Y}s_n(\mathsf{X}))) = \inf_{s\in\mathcal{S}} \mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))) \tag{26}$$

With such a sequence in hand, note that there exists a value $\bar{\theta} > 0$ such that for all integers $n \geq 1$, we have

$$0 \leq \theta^*(s_n) \leq \bar{\theta}. \tag{27}$$

The lower bound in (27) holds because $\phi(\cdot) \geq 0$, and thus the support of the surrogate losses $\phi(\mathsf{Y}s(\mathsf{X}))$ is bounded below by zero.[3] We cannot however guarantee that the support of $\phi(\mathsf{Y}s(\mathsf{X}))$ is bounded above. That said, $(s_1, s_2, \ldots)$ is not a completely arbitrary sequence; from (26) we are assuming this sequence minimizes the OCE criterion. If the sequence $\theta^*(s_n)$ were to grow without bound, it would contradict (26).[4] As such, while the choice of $\bar{\theta} < \infty$ may depend on the sequence, (26) always implies (27).

Next, we would like to link up surrogate losses (under the OCE criterion) to the error probability $\mathcal{E}(\cdot)$. To do this, first note that the excess value in the OCE criterion can be bounded below by

$$\begin{aligned}\mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))) - \inf_{s\in\mathcal{S}} \mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))) &= \mathrm{C}_\phi(s;\theta^*(s)) - \inf_{s'\in\mathcal{S}} \left[\inf_{\theta\in\mathbb{R}} \mathrm{C}_\phi(s';\theta)\right] \\ &\geq \mathrm{C}_\phi(s;\theta^*(s)) - \inf_{s'\in\mathcal{S}} \mathrm{C}_\phi(s';\theta^*(s)).\end{aligned} \tag{28}$$

Next let us note that the normalization of ρ (slope 1, value 0 at the origin) combined with its convexity and monotonicity immediately implies $\rho(u) \geq u$ for all $u \in \mathbb{R}$, and thus for any choice of $\theta \in \mathbb{R}$, we have $\widetilde{\phi}(u;\theta) \geq \theta + (\phi(u) - \theta) = \phi(u)$. As such, it follows that

$$\phi(\cdot) \geq 0 \implies \widetilde{\phi}(\cdot;\theta) \geq 0. \tag{29}$$

This fact is useful because it allows us to leverage structural results of Bartlett et al. (2006). To do so, let us denote the so-called "conditional $\widetilde{\phi}$-risk" by

$$C_\beta(u;\theta) \coloneqq \beta\widetilde{\phi}(u;\theta) + (1-\beta)\widetilde{\phi}(-u;\theta) \tag{30}$$

for each $u \in \mathbb{R}$, $\theta \in \mathbb{R}$, and $0 \leq \beta \leq 1$. Note that we can connect this function to our modified penalty through the equality

$$\mathrm{C}_\phi(s;\theta) = \mathbf{E}_{\mathsf{X}}\left[C_{\beta(\mathsf{X})}(s(\mathsf{X});\theta)\right] \tag{31}$$

[3] That this implies $0 \leq \theta^*(s)$ for any $s \in \mathcal{S}$ is a basic property of OCE criteria; see for example Ben-Tal & Teboulle (2007, Prop. 2.1).

[4] To see this, just note that the θ term in the OCE criterion increases (with θ) no slower than the $\rho(\cdot - \theta)$ term decreases (again, as θ grows), due to the slope, convexity, and monotonicity conditions on ρ.

that holds for any choice of $\theta \in \mathbb{R}$ and $s \in \mathcal{S}$, where we have denoted $\beta(x) \coloneqq \mathbf{P}\{\mathsf{Y} = 1 \,|\, \mathsf{X} = x\}$. Optimal values of this function (with and without constraints) are denoted by

$$H^-(\beta;\theta) \coloneqq \inf\left\{C_\beta(u;\theta) : \, u(2\beta - 1) \leq 0\right\}, \tag{32}$$

$$H(\beta;\theta) \coloneqq \inf_{u\in\mathbb{R}} C_\beta(u;\theta). \tag{33}$$

Here $H(\beta;\theta)$ corresponds to the optimal conditional risk computed in terms of $\widetilde{\phi}$, and $H^-(\beta;\theta)$ is used to quantify the best possible risk when a score (i.e., with u as $s(x)$) disagrees with the Bayes optimal score (i.e., $(2\beta - 1)$ with β as $\mathbf{P}\{\mathsf{Y} = 1 \,|\, \mathsf{X} = x\}$). The so-called "$\psi$-transform" is based on the gap between these two optimal values. Denoting

$$\widetilde{\psi}(u;\theta) \coloneqq H^-\left(\frac{1+u}{2};\theta\right) - H\left(\frac{1+u}{2};\theta\right) \tag{34}$$

for all $-1 \leq u \leq 1$, define $\psi(\cdot;\theta) \coloneqq \widetilde{\psi}^{**}(\cdot;\theta)$, namely the Fenchel-Legendre biconjugate of $\widetilde{\psi}$. The first key structural results we wish to leverage are as follows:

$$\psi(\cdot;\theta) \text{ is continuous on } [0,1], \quad \psi(\cdot;\theta) \geq 0, \quad \psi(0;\theta) = 0. \tag{35}$$

The properties in (35) hold for all $\theta \in \mathbb{R}$.[5] Furthermore, using (29) and applying the proof of Bartlett et al. (2006, Thm. 1(1)), for any choice of $s \in \mathcal{S}$ and $\theta \in \mathbb{R}$, we have

$$\psi\left(\mathcal{E}(h_s) - \mathcal{E}^*;\theta\right) \leq \mathbf{E}_{\mathsf{X}}\left[C_{\beta(\mathsf{X})}(s(\mathsf{X});\theta) - H\left(\beta(\mathsf{X});\theta\right)\right], \tag{36}$$

where $h_s(\cdot) \coloneqq \operatorname{sign}(s(\cdot))$, $\mathcal{E}^* \coloneqq \min_{h\in\mathcal{H}} \mathcal{E}(h)$ and $\beta(x) \coloneqq \mathbf{P}\{\mathsf{Y} = 1 \,|\, \mathsf{X} = x\}$ as before. By assuming that $\mathcal{S}$ is the set of all measurable functions on $\mathcal{X}$, we have

$$H\left(\beta(\mathsf{X});\theta\right) = \inf_{s'\in\mathcal{S}} \mathbf{E}_{\mathsf{X}}\left[C_{\beta(\mathsf{X})}(s'(\mathsf{X});\theta)\right] \tag{37}$$

for any choice of $\theta \in \mathbb{R}$.[6] Combining (36), (37), and (31) with the inequality (28) established earlier, we may conclude that

$$\mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))) - \inf_{s\in\mathcal{S}} \mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))) \geq \psi\left(\mathcal{E}(h_s) - \mathcal{E}^*;\theta^*(s)\right) \tag{38}$$

holds for any choice of $s \in \mathcal{S}$. Applying this bound (38) to any sequence $(s_1, s_2, \ldots)$ satisfying (26), it follows from the basic properties of (35) that

$$\lim_{n\to\infty} \psi\left(\mathcal{E}(h_n) - \mathcal{E}^*;\theta^*(s_n)\right) = 0, \tag{39}$$

where we have denoted $h_n(\cdot) \coloneqq \operatorname{sign}(s_n(\cdot))$ for readability. All that remains is to analyze how the sequence $(\mathcal{E}(h_1), \mathcal{E}(h_2), \ldots)$ behaves.

In order to establish $\mathcal{E}(h_n) \to \mathcal{E}^*$, we need to utilize some additional properties of ρ. We have assumed that ρ is differentiable, and can easily confirm that the first two derivatives of $\widetilde{\phi}(\cdot;\theta)$ are as follows:

$$\widetilde{\phi}'(u;\theta) = \rho'(\phi(u) - \theta)\phi'(u),$$
$$\widetilde{\phi}''(u;\theta) = \rho''(\phi(u) - \theta)(\phi'(u))^2 + \rho'(\phi(u) - \theta)\phi''(u).$$

Since ρ and ϕ are both assumed to be convex, and strict monotonicity of ρ implies $\rho'(\cdot) > 0$, we can immediately conclude that $\widetilde{\phi}''(\cdot;\theta) \geq 0$, and thus $\widetilde{\phi}(\cdot;\theta)$ is convex on $\mathbb{R}$, for any choice of $\theta \in \mathbb{R}$. Furthermore, since we have assumed $\phi'(0) < 0$, it also follows that $\widetilde{\phi}'(0;\theta) < 0$. This property along with convexity implies that we have

$$\psi(u;\theta) > 0, \quad 0 < u \leq 1 \tag{40}$$

$$\psi(u;\theta) = \widetilde{\phi}(0;\theta) - H\left(\frac{1+u}{2};\theta\right) \tag{41}$$

[5] These properties follow from Lemma 2 (parts 6, 7, and 8) of Bartlett et al. (2006) after applying (29).

[6] This fact is also used in the proof of Bartlett et al. (2006, Thm. 1(1)).

for any $\theta \in \mathbb{R}$.[7] Using the definitions of $\widetilde{\phi}$ and $H(\cdot;\theta)$ along with (41), we can write

$$\psi(u;\theta) = \rho(\phi(0)-\theta) - \inf_{v\in\mathbb{R}} \left[\left(\frac{1+u}{2}\right)\rho(\phi(v)-\theta) + \left(\frac{1-u}{2}\right)\rho(\phi(-v)-\theta) \right].$$

Introducing the helper function

$$g(\theta,u;v) \coloneqq \left(\frac{1+u}{2}\right)\rho(\phi(v)-\theta) + \left(\frac{1-u}{2}\right)\rho(\phi(-v)-\theta),$$

clearly we have $\psi(u;\theta) = \rho(\phi(0)-\theta) - \inf_{v\in\mathbb{R}} g(\theta,u;v)$. Due to the monotonicity of ρ and the convexity of both ϕ and ρ, it is easy to directly verify that $(\theta,v) \mapsto g(\theta,u;v)$ is convex on $\mathbb{R}^2$, for any choice of $0 \le u \le 1$. In addition, partial minimization preserves convexity, so $\theta \mapsto \inf_{v\in\mathbb{R}} g(\theta,u;v)$ is also convex and continuous.[8] As such, for any convergent sequence $(\theta_1,\theta_2,\ldots)$ and any choice of $0 \le u \le 1$, we have

$$\lim_{n\to\infty} \psi(u;\theta_n) = \psi\left(u; \lim_{n\to\infty}\theta_n\right). \tag{42}$$

With this fact established, let us say that under (26), the sequence $\mathcal{E}(h_n)$ does not converge to $\mathcal{E}^*$. In such a case, there exists some $\varepsilon > 0$ such that

$$\limsup_{n\to\infty} [\mathcal{E}(h_n) - \mathcal{E}^*] = \varepsilon.$$

From the definition of $\limsup$, we can always find a convergent subsequence, denoted $(s_1', s_2', \ldots)$ such that $\mathcal{E}(h_n') - \mathcal{E}^* \to \varepsilon$, where $h_n'(\cdot) \coloneqq \operatorname{sign}(s_n'(\cdot))$. This implies that for all large enough n, we have $\mathcal{E}(h_n') - \mathcal{E}^* \ge \varepsilon/2 > 0$, and further that we have

$$\psi(\mathcal{E}(h_n') - \mathcal{E}^*; \theta^*(s_n')) \ge \psi(\varepsilon/2; \theta^*(s_n')) > 0 \tag{43}$$

for all such n.[9] Noting that from (27), all elements of sequence $\theta^*(s_n')$ must be included in a bounded interval (i.e., a compact set), there exists a sub-sequence of (s_n') that we denote as $(s_1'', s_2'', \ldots)$, such that $\theta^*(s_n'')$ converges.[10] Utilizing the continuity property (42) and positivity (40), it follows that

$$\lim_{n\to\infty} \psi(\varepsilon/2; \theta^*(s_n'')) = \psi\left(\varepsilon/2; \lim_{n\to\infty}\theta^*(s_n'')\right) > 0. \tag{44}$$

This however would contradict (39), which we have already shown to be implied by any sequence satisfying (26). As such, we can conclude that under (26), we have $\mathcal{E}(h_n) \to \mathcal{E}^*$.

Finally, for our statement regarding the inclusion of solution sets, for the trivial case that no minimizer of $h \mapsto \mathrm{OCE}(\mathsf{L}_\phi(h))$ exists, we have $\emptyset \subset \mathcal{H}_{\mathcal{E}}^*$; let us consider the non-trivial case where a minimizer exists. Write such a minimizer by

$$s^* \in \operatorname*{arg\,min}_{s\in\mathcal{S}} \mathrm{OCE}(\phi(\mathsf{Y}s(\mathsf{X}))). \tag{45}$$

Writing $h^*(\cdot) \coloneqq \operatorname{sign}(s^*(\cdot))$, if $\mathcal{E}(h^*) > \mathcal{E}^*$ were to hold, then via positivity (40), we would have a contradiction to the inequality (38). As such, $h^* \in \arg\min_{h\in\mathcal{H}} \mathcal{E}(h)$ must hold, concluding the proof. □

Proof of Proposition 3.4. To start, fix a threshold $\theta = \theta_\varepsilon \coloneqq \phi(0) + \varepsilon$ using some $\varepsilon > 0$ to be determined later. Next, note that for any $h \in \mathcal{H}$, the expected surrogate loss can be written as

$$\mathbf{E}[\mathsf{L}_\phi(h)] = \mathbf{E}[\mathrm{I}\{h(\mathsf{X}) \neq \mathsf{Y}\}\,\mathsf{L}_\phi(h)] + \mathbf{E}[\mathrm{I}\{h(\mathsf{X}) = \mathsf{Y}\}\,\mathsf{L}_\phi(h)]. \tag{46}$$

[7] The properties (40) and (41) follow by applying Theorem 2 and Lemma 2(9) of Bartlett et al. (2006).

[8] See for example Bertsekas (2015, B.3.3 and B.1.3).

[9] The first inequality holds using the fact that monotonicity of $u \mapsto \psi(u;\theta)$ is implied by the properties of being minimized at zero (see (35)) and convexity on $[0,1]$ for all $\theta \in \mathbb{R}$; see Lemma 2(2) of Bartlett et al. (2006). The second inequality follows from (41).

[10] See for example Rudin (1976, Thm. 2.37).

Now, taking an arbitrary $h^* \in \mathcal{H}^*_{\mathcal{E}}$ and the assumption that $\mathcal{E}(h^*) = 0$, it follows that

$$\mathbf{E}[\mathrm{I}\{h^*(\mathsf{X}) \neq \mathsf{Y}\}\, \mathsf{L}_\phi(h^*)] = 0. \tag{47}$$

Combining (46) and (47) with the monotonicity of ϕ, we have that the expected surrogate loss of $h^* \in \mathcal{H}^*_{\mathcal{E}}$ can be bounded above by

$$\mathbf{E}[\mathsf{L}_\phi(h^*)] = \mathbf{E}[\mathrm{I}\{h^*(\mathsf{X}) = \mathsf{Y}\}\, \mathsf{L}_\phi(h^*)] \leq \phi(0). \tag{48}$$

With this in place, consider any $h \notin \mathcal{H}^*_{\mathcal{E}}$, namely any $h \in \mathcal{H}$ such that $\mathcal{E}(h) > 0$. Recall that by definition of $\mathcal{H}$, there is some $s \in \mathcal{S}$ such that $h(x) = \mathrm{sign}(s(x))$ for all $x \in \mathcal{X}$. Again by assumption, this $s(\cdot)$ is measurable.[11] Consider a simple re-scaling of s, namely for some constant $b > 0$, let us define

$$\widetilde{s}(x) := \begin{cases} b, & s(x) \geq 0 \\ -b, & s(x) < 0. \end{cases}$$

Using the measurability of $s(\cdot)$ and the fact that $\widetilde{s}(\cdot)$ is a simple function (in the usual measure theoretical sense), we have that $\widetilde{s} \in \mathcal{S}$ and thus defining $\widetilde{h}(x) := \mathrm{sign}(\widetilde{s}(x))$, we have $\widetilde{h} \in \mathcal{H}$ as well. In fact, since signs are not changed, we have $h(x) = \widetilde{h}(x)$ for all $x \in \mathcal{X}$, and thus $\mathcal{E}(\widetilde{h}) = \mathcal{E}(h) > 0$. On the other hand, re-scaling impacts the margin penalties incurred via the surrogate loss, and the resulting expected value is

$$\begin{aligned} \mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] &= \mathbf{E}[\mathrm{I}\{\widetilde{h}(\mathsf{X}) \neq \mathsf{Y}\}\, \mathsf{L}_\phi(\widetilde{h})] + \mathbf{E}[\mathrm{I}\{\widetilde{h}(\mathsf{X}) = \mathsf{Y}\}\, \mathsf{L}_\phi(\widetilde{h})] \\ &= \mathcal{E}(h)\phi(-b) + (1 - \mathcal{E}(h))\phi(b). \end{aligned} \tag{49}$$

Using the running assumption that $\phi(\cdot)$ is both unbounded and non-negative, from (49) it is clear that we can take $b > 0$ large enough that $\mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] > \phi(0)$. Returning to the threshold θ_ε, if we set $\varepsilon = \mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] - \phi(0)$, we have found a value of θ_ε such that

$$\mathbf{E}[\mathsf{L}_\phi(h^*)] \leq \phi(0) < \mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] = \theta_\varepsilon,$$

noting that the first inequality comes from (48). Note that this implies

$$\mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] - \theta_\varepsilon = 0 > \mathbf{E}[\mathsf{L}_\phi(h^*)] - \theta_\varepsilon$$

and using the local strict convexity of $\widetilde{\rho}$ near the origin, we have

$$\widetilde{\rho}(\mathbf{E}[\mathsf{L}_\phi(\widetilde{h})] - \theta_\varepsilon) = 0 < \widetilde{\rho}(\mathbf{E}[\mathsf{L}_\phi(h^*)] - \theta_\varepsilon).$$

In other words, we have that $h^* \in \mathcal{H}^*_{\mathcal{E}}$ is sub-optimal in terms of $\mathrm{C}_{\text{inn}}(\mathsf{L}_\phi(\cdot); \theta_\varepsilon)$. Note that the preceding argument (in particular the choice of b) does not depend on the particular choice of $h^* \in \mathcal{H}^*_{\mathcal{E}}$, and thus we can conclude that all elements of $\mathcal{H}^*_{\mathcal{E}}$ share this sub-optimality, which is the desired result for $\mathrm{C} = \mathrm{C}_{\text{inn}}$.

As for the remaining case of $\mathrm{C} = \mathrm{C}_{\text{out}}$, the preceding argument can fail because a large b used in defining $\widetilde{s}$ means that the distribution of $\mathsf{L}_\phi(\widetilde{h})$ becomes arbitrarily widely spread out, and thus can be strongly penalized by C_{out}. As a simple alternative, we consider re-scaling and flipping the sign of any $\mathcal{E}(\cdot)$-optimal classifier. Take any $h^* \in \mathcal{H}^*_{\mathcal{E}}$, with $h^*(x) = \mathrm{sign}(s^*(x))$ for some $s^* \in \mathcal{S}$. Define a modified version of the scoring function as follows:

$$\widetilde{s}^*(x) := \begin{cases} \frac{-s^*(x)}{|s^*(x)|}, & s^*(x) \neq 0 \\ 0, & s^*(x) = 0. \end{cases}$$

As mentioned earlier, measurability is preserved, and so $\widetilde{s}^* \in \mathcal{S}$ and thus setting $\widetilde{h}^*(x) := \mathrm{sign}(\widetilde{s}^*(x))$, we also have $\widetilde{h}^* \in \mathcal{H}$. Since this classifier always disagrees with h^*, we have $1 = \mathcal{E}(\widetilde{h}^*) > \mathcal{E}(h^*) = 0$. With probability 1, we have

$$0 \leq \mathsf{L}_\phi(h^*) \leq \phi(0) < \mathsf{L}_\phi(\widetilde{h}^*) = \phi(-1).$$

[11] We have a typical notion of Borel measurability in mind. It is sufficient, for example, if for any $a \in \mathbb{R}$, we have that the event $\{s(\mathsf{X}) \leq a\}$ is measurable, i.e., an element of the sigma-field composing the underlying measure space. See for example Ash & Doléans-Dade (2000, Ch. 1) for more background.

As such, by taking $\theta = \phi(-1)$, clearly we have that

$$\mathbf{E}[\widetilde{\rho}(\mathsf{L}_\phi(\widetilde{h}^*) - \theta)] = 0 < \mathbf{E}[\widetilde{\rho}(\mathsf{L}_\phi(h^*) - \theta)],$$

where the inequality holds because $\mathsf{L}_\phi(h^*) - \phi(-1) < 0$ holds with probability 1, plus the strict convexity of $\widetilde{\rho}$ around the origin. This shows how while h^* is optimal in $\mathcal{E}(\cdot)$, it is sub-optimal in terms of $\mathrm{C}_{\text{out}}(\mathsf{L}_\phi(\cdot); \theta)$. Again, this approach holds for any choice of $h^* \in \mathcal{H}^*_{\mathcal{E}}$, and thus the desired result holds for $\mathrm{C} = \mathrm{C}_{\text{inn}}$ as well. □

B.2. Divergence of CVaR from error probability

As mentioned in Remark 2.6, CVaR satisfies the non-decreasing condition of Proposition 2.5, but $\rho(u) = (u)_+/(1-\beta)$ is not strictly monotonic on $\mathbb{R}$ (i.e., not increasing). With ρ that is non-decreasing, we know the inclusion (20) holds, telling us that all error probability minimizers are also optimal in terms of CVaR. Are there any CVaR-optimal solutions that are *not* optimal in terms of error probability? The answer is that it depends on both $\mathcal{H}$ and β. To consider this a bit more precisely, first note that quantiles (recall §2.3.2) of the zero-one loss are also binary valued, namely that we have

$$\mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = \begin{cases} 1, & \text{if } 1 > \beta > \mathbf{P}\{\mathsf{L}_{01}(h) = 0\} \\ 0, & \text{if } 0 < \beta \leq \mathbf{P}\{\mathsf{L}_{01}(h) = 0\}. \end{cases}$$

It is straightforward to check that the following implications hold:

$$\min_{h \in \mathcal{H}} \mathcal{E}(h) > 1 - \beta \implies \mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 1, \text{ for all } h \in \mathcal{H} \tag{50}$$

$$\max_{h \in \mathcal{H}} \mathcal{E}(h) \leq 1 - \beta \implies \mathrm{Q}_\beta(\mathsf{L}_{01}(h)) = 0, \text{ for all } h \in \mathcal{H} \tag{51}$$

When the model $\mathcal{H}$ is poor enough (or β is large/strict enough) that the condition in (50) holds, the CVaR risk is constant, i.e., $\underline{\mathrm{C}}_\rho(\mathsf{L}_{01}(h)) = 1$ for all $h \in \mathcal{H}$, and thus trivially all $h \in \mathcal{H}$ are optimal in CVaR, but certainly need not be in $\mathcal{E}(\cdot)$. When the model is good enough (or β is small/loose enough) that the condition in (51) holds, we have for all $h \in \mathcal{H}$ that

$$\rho(1 - \mathrm{Q}_\beta(\mathsf{L}_{01}(h))) - \rho(-\mathrm{Q}_\beta(\mathsf{L}_{01}(h))) = \rho(1) > 0,$$

and thus the strong monotonicity argument used earlier applies, telling us that CVaR-optimal solutions and error probability minimizers are identical (i.e., (21) holds).

B.3. Collapse under variantile

With Remark 2.9 as context, note that for $0 \leq \theta \leq 1$, it is straightforward to confirm that under the zero-one loss $\mathsf{L} = \mathsf{L}_{01}$, for any $h \in \mathcal{H}$ we have

$$2\mathbf{E}\left[|\mathbf{1}_\theta(\mathsf{L}_{01}(h)) - \tau|(\mathsf{L}_{01}(h) - \theta)^2\right] = 2\left[(1-\tau)(1-\mathcal{E}(h))\theta^2 + \tau\mathcal{E}(h)(1-\theta)^2\right].$$

Checking first-order conditions, this function can be minimized (in θ) by setting

$$\theta = \frac{\tau\mathcal{E}(h)}{(1-\tau)(1-\mathcal{E}(h)) + \tau\mathcal{E}(h)}.$$

Plugging this solution in to obtain an explicit form of $\underline{\mathrm{C}}_\tau(\mathsf{L}_{01}(h))$, we have

$$\underline{\mathrm{C}}_\tau(\mathsf{L}_{01}(h)) = \\ 2\left[(1-\tau)(1-\mathcal{E}(h))\left(\frac{\tau\mathcal{E}(h)}{(1-\tau)(1-\mathcal{E}(h)) + \tau\mathcal{E}(h)}\right)^2 + \tau\mathcal{E}(h)\left(\frac{(1-\tau)(1-\mathcal{E}(h))}{(1-\tau)(1-\mathcal{E}(h)) + \tau\mathcal{E}(h)}\right)^2\right].$$

With this form in mind, simply replace $\mathcal{E}(h)$ with a variable p that can be taken over $[0,1]$, and introduce a helper function

$$g_\tau(p) \coloneqq 2\left[(1-\tau)(1-p)\left(\frac{\tau p}{(1-\tau)(1-p) + \tau p}\right)^2 + \tau p\left(\frac{(1-\tau)(1-p)}{(1-\tau)(1-p) + \tau p}\right)^2\right]. \tag{52}$$

One can readily check that this function $g_\tau(\cdot)$ is strictly concave on the unit interval for any choice of $0 < \tau < 1$; see Figure 3 for a numerical example. As a result, we may conclude that depending on the nature of $\mathcal{H}$, the minimizer set of $\underline{\mathrm{C}}_\tau(\mathsf{L}_{01}(\cdot))$ is either the set of minimizers *or* maximizers of $\mathcal{E}(\cdot)$, a result analogous to the point raised in §2.3.1.

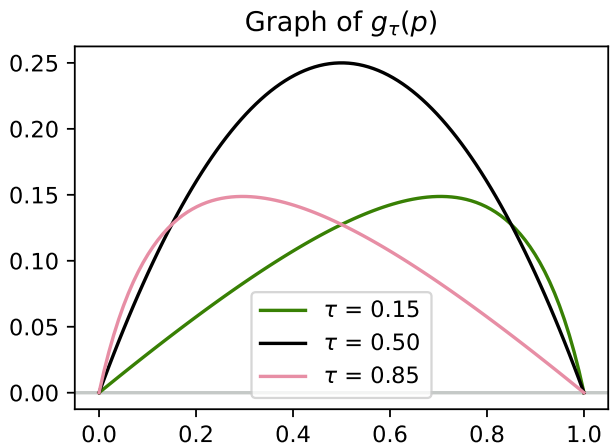

Figure 3. Graphs of $g_\tau(\cdot)$ in (52) over the unit interval for varying choices of τ.

C. Additional Figures

As supplementary material for the main text, here we include two additional figures.

- Examples of valid ρ for OCE criteria and $\widetilde{\rho}$ for loss-restraining criteria are shown in Figure 4.
- In §4 we only gave metric trajectories for CIFAR-100 and SVHN; analogous results for CIFAR-10 and FashionMNIST are given in Figure 5.

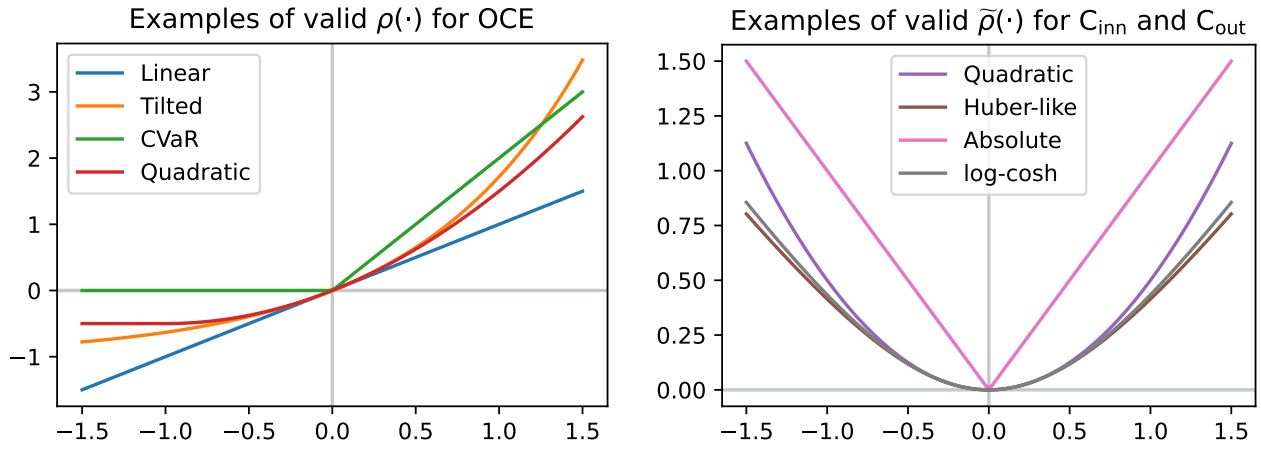

Figure 4. Examples of valid choices of ρ (left) and $\widetilde{\rho}$ (right) for use in defining OCE criteria (13) and loss-restraining criteria (14)–(15) respectively.

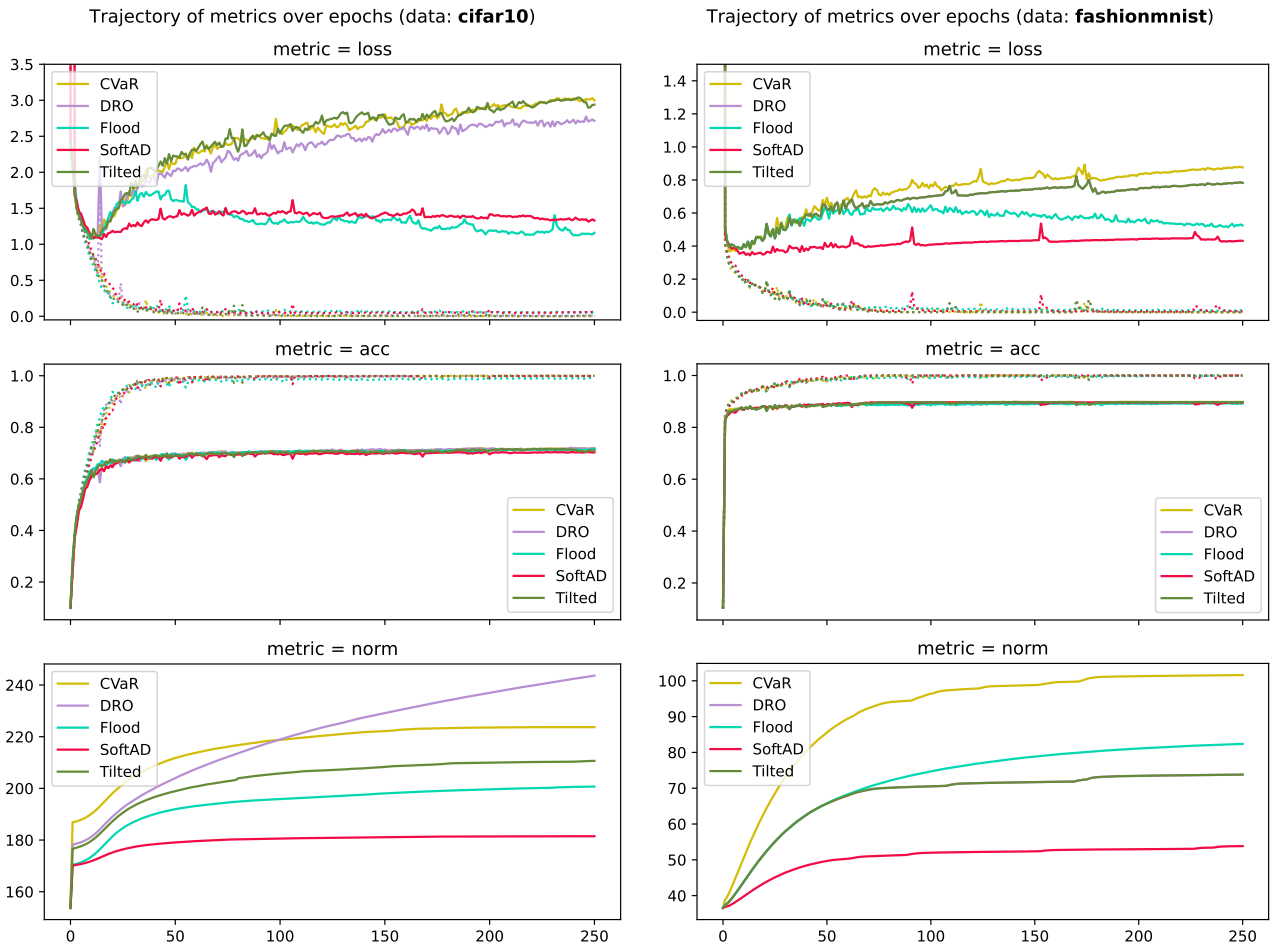

Figure 5. Results for CIFAR-10 and FashionMNIST; see the caption of Figure 2 for details.