Distribution Free M-estimation

Felipe Areces **John Duchi**
Stanford University

Abstract

The basic question of delineating those statistical problems that are solvable without making any assumptions on the underlying data distribution has long animated statistics and learning theory. This paper characterizes when a convex M-estimation or stochastic optimization problem is solvable in such an assumption-free setting, providing a precise dividing line between solvable and unsolvable problems. The conditions we identify show that Lipschitz continuity of the loss being minimized is not necessary for distribution free minimization, and they are also distinct from classical characterizations of learnability in machine learning.

1 INTRODUCTION

Consider a general convex M-estimation problem, where for a population distribution P on a set $\mathcal{Z}$ and a convex loss $\ell_z(\theta)$ measuring the performance of a parameter θ on example z, we wish to minimize the population loss

$$L_P(\theta) := \mathbb{E}_P[\ell_Z(\theta)] = \int \ell_z(\theta) dP(z) \tag{1}$$

over θ belonging to a convex parameter space Θ in Euclidean space. We are interested in truly distribution free minimization of this population loss, meaning that we would like to be able to (asymptotically) minimize L_P given i.i.d. observations Z_i drawn from P without making *any* assumptions on the distribution P. By this, we mean we know essentially the bare minimum from a statistical perspective: only (i) the loss ℓ, (ii) the set $\mathcal{Z}$, and (iii) the parameter space Θ. Our main contribution will be to delineate those situations in which minimizing the loss (1) is possible from those in which it is not.

To do so, we require a bit more formality. For each $z \in \mathcal{Z}$, the loss $\ell_z : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed convex and proper, and to avoid trivialities, we assume that $\ell_z(\theta) < \infty$ for all $\theta \in \operatorname{int}\Theta$ and is measurable in z. Defining the minimal loss $L_P^\star(\Theta) := \inf_{\theta\in\Theta} L_P(\theta)$ and letting $\mathcal{P}(\mathcal{Z})$ be the collection of (Borel) probability measures on $\mathcal{Z}$, we give conditions under which the *minimax optimization risk* for the loss ℓ,

$$\mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) := \inf_{\widehat{\theta}_n} \sup_{P\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{P^n}\Big[L_P(\widehat{\theta}_n(Z_1^n)) - L_P^\star(\Theta)\Big], \tag{2}$$

approaches 0 as n grows—meaning M-estimation (optimization) is possible—or, conversely, is bounded away from 0. The definition (2) takes an infimum over measurable estimators $\widehat{\theta}_n : \mathcal{Z}^n \to \Theta$ and supremum over distributions P generating $Z_1^n = (Z_1, \ldots, Z_n) \stackrel{\text{iid}}{\sim} P$. Crucially, in the minimax risk (2), the loss ℓ and sets $\mathcal{Z}$ and Θ are fixed, so any lower bound holds not because we may construct a worst-case function but because of properties the instance at hand actually enforces. This makes clear the sense in which we consider an estimator distribution free: it must achieve small excess risk $L_P(\widehat{\theta}) - L_P^\star$ uniformly over sampling distributions P for a given triple $(\ell, \mathcal{Z}, \Theta)$.

If the loss ℓ_z is Lipschitz for each $z \in \mathcal{Z}$ with the same Lipschitz constant M, then stochastic gradient algorithms (Nemirovski et al., 2009) achieve

$$\mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) \le \frac{M \operatorname{diam}(\Theta)}{\sqrt{n}}. \tag{3}$$

These guarantees are tight in that there exist sets $\mathcal{Z}$ and M-Lipschitz convex losses ℓ for which $\mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) \ge c\frac{M \operatorname{diam}(\Theta)}{\sqrt{n}}$, where $c > 0$ is a numerical constant (Raginsky and Rakhlin, 2011; Agarwal et al., 2012). In contrast to these "classic" lower bounds, which find worst-case losses and sets $\mathcal{Z}$ to demonstrate tight convergence results, we provide a precise dividing line, based on the properties of the particular loss ℓ and its behavior on the space $\mathcal{Z} \times \Theta$, for the separation

$$\lim_n \mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) > 0 \quad \text{versus} \quad \lim_n \mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) = 0.$$

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

(In the generality we consider here, there is no hope of getting a convergence rate, as we will show.)

We will give the dividing lines separating these cases both when Θ is compact and when it is unbounded. In the compact case, our main results show that the following condition very nearly provides this division:

Condition 1: For each compact subset $\Theta_0 \subset \text{int}\,\Theta$, the functions $\ell_z(\cdot)$ restricted to Θ_0 are uniformly Lipschitz: there exists $M = M(\Theta_0) < \infty$ such that for each $z \in \mathcal{Z}$, the function $\ell_z(\cdot)$ is M-Lipschitz continuous on Θ_0.

If the condition holds, then $\lim_n \mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) = 0$: there exist estimators that can solve the minimization problem with vanishing risk uniformly over all P. If it fails, then excepting some trivialities about achievable minimizers that we elucidate, no estimation or optimization procedure can achieve excess risk tending to zero uniformly over distributions P on $\mathcal{Z}$.

Questions of truly distribution free inference date back at least to Bahadur and Savage's work on the impossibility of nonparametric inference of a mean (1956). A thread of such work stretches through today; for example, Donoho (1988) shows that two-sided confidence statements, i.e., of the form "with 95% confidence, the underlying distribution P has KL-divergence from the uniform between 1 and 3" are generally impossible. More recently, *conformal prediction* methods perform predictive inference without making any assumptions on the underlying distribution (e.g. Vovk et al., 2005; Vovk, 2013). At the most basic level, conformal prediction methods leverage the ability to robustly estimate quantiles to provide (marginal) confidence statements on the predictions of even completely black-box models (Lei, 2014; Lei and Wasserman, 2014; Lei et al., 2018; Barber et al., 2021). This renewed focus on assumption free inference motivates the questions we investigate in M-estimation.

Work on *universal consistency* in nonparametric regression, beginning with Cover and Hart (1967), asks related questions to ours. There, the data is $z = (x, y)$, and one wishes to estimate $f^\star(x) := \mathbb{E}[Y \mid X = x]$. An estimator $\widehat{f}_n$ based on $(X_i, Y_i)_{i=1}^n \stackrel{\text{iid}}{\sim} P$ is *consistent* if $\mathbb{E}[|\widehat{f}_n(X_{n+1}) - f^\star(X_{n+1})|] \to 0$ as $n \to \infty$ whenever $\mathbb{E}[|Y|] < \infty$. Cover and Hart (1967) and Stone (1977) show that k-nearest neighbor classifiers, where $k = k_n \to \infty$ but $k_n/n \to 0$, are consistent for (respectively) classification and regression. In distinction from our setting, however, these results are pointwise (not uniform in the distribution P), as uniform guarantees are impossible without making additional assumptions on the regression function $f^\star$. See also the books of Devroye et al. (1996); Györfi et al. (2002).

In theoretical machine learning, researchers studying *learnability* investigate problems similar to ours (Vapnik, 1998). For prediction problems with data $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, one seeks a hypothesis $h : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ to minimize $L_P(h) := \mathbb{E}_P[\ell(h(X), Y)]$ for a loss ℓ. A problem is *learnable* for the class $\mathcal{H}$ if

$$\inf_{\widehat{h}} \sup_P \mathbb{E}_P \left[L_P(\widehat{h}_{P_n}) - \inf_{h \in \mathcal{H}} L_P(h) \right] \to 0, \qquad (4)$$

where $\widehat{h}_{P_n}$ is a function of the empirical distribution P_n; this coincides with convergence of the minimax risk (2). Early work considered particular losses ℓ; in the case of binary classification, where $\mathcal{Y} = \{\pm 1\}$ and $\ell(\hat{y}, y) = 1\{\hat{y} \neq y\}$, learnability (2) holds if and only if $\mathcal{H}$ has finite VC-dimension, which in turn coincides with the uniform convergence that $\sup_{h \in \mathcal{H}} |L_{P_n}(h) - L_P(h)| \to 0$ (see Anthony and Bartlet (1999, Ch. 19) or Alon et al. (1997)). Shalev-Shwartz et al. (2010) generalize these results to show that when the loss $\ell(h(x), y)$ is uniformly bounded, a class $\mathcal{H}$ is learnable (4) if and only if there exist leave-one-out stable procedures asymptotically performing empirical risk minimization, i.e., $\mathbb{E}[L_{P_n}(\widehat{h}_{P_n}) - \inf_{h \in \mathcal{H}} L_{P_n}(h)] \to 0$ uniformly in P.

Condition 1 gives distinct conditions from this prior work. We dedicate the remainder of the paper to demonstrating the ways in which Condition 1 describes when M-estimation is possible and impossible. Section 2 provides the main results, with Sec. 2.2 addressing the case when Θ is a compact set. In Section 2.3, we present Theorem 3, which precisely characterizes when the minimax risk for M-estimation and optimization can tend to zero when Θ may be noncompact (providing a Condition 2). Perhaps surprisingly, given the upper and lower bounds using Lipschitz constants (3), these results make clear that global Lipschitz continuity of the losses ℓ_z is neither sufficient nor necessary for distribution-free minimization.

The negative results in Theorem 3 apply to traditional cases of "robust" estimation, such as the absolute loss $\ell_z(\theta) = |\theta - z|$, corresponding to estimating a median. Thus, we briefly consider the problem of finding stationary points of L_P instead of points for which the excess loss $L_P(\theta) - L_P^\star$ is small in Section 3, which relates more closely to questions of quantile prediction. We do not characterize the separations in minimax risk, instead giving a convergence result that applies to both differentiable and non-differentiable functions, showing that M-estimation, as we have defined it, necessarily differs from obtaining stationary points. Our discussion (see Section 4) builds out of these differences, providing open questions and identifying alternatives to the choice (2) of optimality criterion.

Notation We use $\mathbf{1}_z$ to denote a point mass at the point z. For a convex function f, $\partial f(\theta_0)$ denotes the subdifferential of f at x, meaning those vectors g for which $f(\theta) \ge f(\theta_0) + \langle g, y - \theta_0 \rangle$ for all θ. We let $\mathbb{B}_2 = \{v \mid \|v\|_2 \le 1\}$ be the ℓ_2 ball (in dimension that is clear from context), while $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ denotes the sphere. We let $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ and $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$ denote the extended up (or down) reals. For any set $A \subset \mathcal{Z}$, we let $\mathcal{P}(A)$ denote the collection of Borel probability distributions on A. We also let $[f]_+ := \max\{f, 0\}$.

2 MINIMAX BOUNDS ON RISK

To make rigorous that Condition 1 neatly divides solvable versus unsolvable problems, we present preliminary definitions to avoid minutiae, then provide formal minimax lower bounds in Sections 2.2 and 2.3.

2.1 Restricting to achievable minimizers

We shall restrict analysis in the minimax lower bounds to what we term *achievable minimizers*, roughly meaning those points in Θ that could plausibly minimize $L_P(\theta) = \mathbb{E}_P[\ell_Z(\theta)]$ for some P on $\mathcal{Z}$. This restricts little; in Appendix A, where we enumerate properties of and provide technical detail on the achievable minimizers, we also see that there are points in the set of achievable minimizers with smaller risk than those outside the set. We also warn the reader that this section simply helps us avoid pathologies; it might be skipped on a first reading.

In the one-dimensional case, the intuition is simple. Assuming temporarily that ℓ is differentiable, we only consider points θ for which there exist z^+ and z^- satisfying $\ell'_{z^+}(\theta) > 0$ and $\ell'_{z^-}(\theta) < 0$. Then immediately, θ minimizes L_P for a distribution $P = p\mathbf{1}_{z^+} + (1-p)\mathbf{1}_{z^-}$ for some appropriate $p \in (0, 1)$: we cannot eliminate the possibility that θ minimizes L_P for some P without observing data. Because derivatives of convex functions are monotonic, if $\ell'_z(\theta) \le 0$ for all $z \in \mathcal{Z}$, then for any distribution P, there *necessarily* exists a minimizer $\theta^\star(P)$ of $L_P(\theta) = \mathbb{E}_P[\ell_Z(\theta)]$ with $\theta^\star(P) \ge \theta$, and similarly when $\ell'_z(\theta) \ge 0$ for all z. We consequently define the minimal and maximal plausible values

$$\begin{aligned} \theta_{\min} = \theta_{\min}(\ell, \mathcal{Z}, \Theta) &:= \inf_{\theta \in \Theta} \left\{ \theta \mid \sup_{z \in \mathcal{Z}} \ell'_z(\theta) > 0 \right\} \\ \theta_{\max} = \theta_{\max}(\ell, \mathcal{Z}, \Theta) &:= \sup_{\theta \in \Theta} \left\{ \theta \mid \inf_{z \in \mathcal{Z}} \ell'_z(\theta) < 0 \right\}, \end{aligned} \tag{5}$$

where if the set defining $\theta_{\min}$ is empty we take the right endpoint $\theta_{\min} = \sup\{\theta \in \Theta\}$ and similarly $\theta_{\max} = \inf\{\theta \in \Theta\}$ in the complementary case. (We take extrema over points for which the $\ell'_z(\theta)$ exists; Appendix A shows the differentiability is immaterial.)

It is possible that $\theta_{\max} \le \theta_{\min}$, but this trivializes the problem, as any $\theta^\star \in [\theta_{\max}, \theta_{\min}]$ minimizes L_P simultaneously for all distributions P; see Lemma A.1 in Appendix A. Thus, define

$$\Theta_{\text{ach}} := [\theta_{\min}, \theta_{\max}], \tag{6}$$

passing tacitly to the extended reals when Θ is unbounded to allow sets of the form $[a, \infty]$ or $[-\infty, b]$ (cf. Hiriart-Urruty and Lemaréchal, 1993a, Appendix A.2). Intuitively, because $L'_P(\theta) \ge 0$ for $\theta > \theta_{\max}$ and $L'_P(\theta) \le 0$ for $\theta < \theta_{\min}$ for *all* P, restricting analysis to the set Θ_{ach} changes nothing.

When $\Theta \subset \mathbb{R}^d$ for general $d \ge 1$, we require a number of technical conditions extending the condition (5). We defer formal definitions to Appendix A.2, instead focusing here on a sufficient condition that mirrors condition (5). Recall the *directional derivative* (Hiriart-Urruty and Lemaréchal, 1993a, Def. VI.1.1.1) of a convex function f at the point $\theta \in \operatorname{dom} f$ in the direction v,

$$\mathsf{D}f(\theta; v) := \lim_{t \downarrow 0} \frac{f(\theta + tv) - f(\theta)}{t} \stackrel{(\star)}{=} \sup_{g \in \partial f(\theta)} \langle g, v \rangle$$

where the limit always exists, $\mathsf{D}f(\theta; v)$ is convex and positively homogeneous in v, and the equality $(\star)$ holds if $\partial f(\theta) \ne \emptyset$. If f is differentiable at θ, then $\mathsf{D}f(\theta; v) = \langle \nabla f(\theta), v \rangle$. We now propose the following definition.

Definition 2.1. *A point θ is* achievable *if for each $v \in \mathbb{S}^{d-1}$,*

$$\textit{there exists } z \in \mathcal{Z} \textit{ with } \mathsf{D}\ell_z(\theta; v) > 0.$$

Typical losses in statistics and machine learning satisfy these achievability conditions; an example may help.

Example 1 (Generalized linear models)**:** Consider a GLM for predicting $y \in \mathcal{Y} \subset \mathbb{R}$ from $x \in \mathcal{X} \subset \mathbb{R}^d$ with $p_\theta(y \mid x) = \exp(y\langle x, \theta \rangle - A(\theta \mid x))h(y)$, where h is a carrier and A is the log-partition function (Brown, 1986; Wainwright and Jordan, 2008). Then because

$$\nabla A(\theta \mid x) = \mathbb{E}_\theta[Y \mid X = x] \cdot x$$

(where $\mathbb{E}_\theta$ denotes expectation under p_θ), the log loss $\ell_{x,y}(\theta)$ is $\mathcal{C}^\infty$ and satisfies $\nabla \ell_{x,y}(\theta) = x(\mathbb{E}_\theta[Y \mid x] - y)$. As $\mathbb{E}_\theta[Y \mid x] \in \operatorname{int} \operatorname{Conv}(\mathcal{Y})$, so long as $\mathcal{X}$ contains *some* linearly independent set of d points, the condition in Definition 2.1 holds for all $\theta \in \mathbb{R}^d$. $\diamond$

When $\Theta \subset \mathbb{R}^d$ for general $d \ge 1$, we require a technical extension to the condition (5):

Definition 2.2. *A set $C \subset \Theta$ is* directable *if for each $\epsilon > 0$, there is a compact convex C_ϵ satisfying $C \subset C_\epsilon \subset \Theta$ and a collection of points $\{z_i\}_{i=1}^k \subset \mathcal{Z}$ such that* *(i)* $\sup_{\theta \in C_\epsilon} \operatorname{dist}(\theta, C) \le \epsilon$, *and*

(ii) there exists $\delta > 0$ *for which*

$$\bigcup_{Q \in \mathcal{P}(\{z_i\}_{i=1}^k)} \bigcup_{\theta \in C_\epsilon} \partial \left(L_Q + \mathbb{I}_\Theta\right)(\theta) \supset \delta \mathbb{B}.$$

Implicit in the Definition 2.2 is that $\partial(L_Q + \mathbb{I}_\Theta)$ is nonempty. In the one-dimensional case, for each $\theta \in \operatorname{int} \Theta$, the point $\{\theta\}$ is directable whenever there are z^+, z^- such that $\ell'_{z^+}(\theta) > 0$ and $\ell'_{z^-}(\theta) < 0$, so that this generalizes the sets defining the extreme plausible values (5). Intuitively, a directable set means that in an arbitrarily small neighborhood of the set, we can "place" a subgradient in any direction $u \in \mathbb{S}^{d-1}$: there exist θ and Q for which $\delta u \in \partial L_Q(\theta)$. Appendix A.2 provides several sufficient characterizations for directability to show that such sets are a typical phenomenon.

We move to the definition of the achievable set in the general d-dimensional case, which we perform by considering the outer construction of closed convex sets as intersections of half-spaces. To that end, for $v \in \mathbb{R}^d$ and $t \in \mathbb{R}$, define the half space $H_{v,t} \coloneqq \{\theta \mid \langle v, \theta \rangle \le t \left\|v\right\|_2\}$, and call the pair $(v, t) \in \mathbb{R}^d \times \mathbb{R}$ *unconstraining* if

$$H_{v,t} \cap C \neq \emptyset \quad \text{for all directable } C \subset \Theta.$$

That is, $H_{v,t}$ has non-trivial intersection with all directable sets C. Call the collection of such pairs $\mathcal{U}$. When Θ is compact, any unconstraining half space $H_{v,t}$ satisfies

$$L_Q^\star(H_{v,t} \cap \Theta) = L_Q^\star(\Theta) \tag{7}$$

for all $Q \in \mathcal{P}_{\rm disc}$, because $\operatorname{argmin}_{\theta \in \Theta} L_Q(\theta)$ is a nonempty compact convex set; see Lemma A.3 in Appendix A.2. In fact, equality (7) extends to any (Borel) probability measure P for which the loss ℓ_z is integrable; see also Lemma A.5 in Appendix A.2. More generally, whenever Q is such that $\operatorname{argmin}_{\theta \in \Theta} L_Q(\theta)$ exists and is compact, the equality (7) holds for all $(v, t) \in \mathcal{U}$.

Noting that $v = \mathbf{0}$ allows $H_{v,t} = \mathbb{R}^d$, we may then define the achievable set

$$\Theta_{\rm ach} \coloneqq \bigcap_{(v,t) \in \mathcal{U}} H_{v,t}. \tag{8}$$

Clearly, $\Theta_{\rm ach} \subset \Theta$ as Θ is closed convex, and in one dimension, the half-spaces $H_{1,t} = (-\infty, t]$ and $H_{-1,t} = [t, \infty)$ show that definitions (6) and (8) coincide; see also Lemma A.2 in Section A.

The next example highlights why, even in statistical or other optimization scenarios, we may wish to consider settings in which the achievable points (Definition 2.1) are not dense.

Example 2 (Robust losses for location)**:** Consider estimating a location via a robust loss $\ell_z(\theta) = \|\theta - z\|$ for $\Theta = \mathbb{R}^d$, where $z \in \mathcal{Z} \subset \mathbb{R}^d$ and $\|\cdot\|$ is a norm on $\mathbb{R}^d$. Even if $\mathcal{Z}$ is discrete (e.g., taking values with integer coordinates), if $\mathcal{Z}$ covers enough space that $\operatorname{Conv}(\mathcal{Z}) = \mathbb{R}^d$, then because $\theta = z$ uniquely minimizes $\ell_z(\theta)$ for each z, the set of unconstraining half-spaces in the definition (8) consists of $\{H_{\mathbf{0},t}\}_{t \in \mathbb{R}}$, so $\Theta_{\rm ach} = \mathbb{R}^d$. $\diamond$

We provide further discussion of $\Theta_{\rm ach}$ and its properties in Appendix A, but for now, content ourselves with noting that $\Theta_{\rm ach}$ contains all achievable points (which, as in the case of GLMs, consists of all points θ). We therefore make the following standing assumption without further comment and with essentially no loss of generality except that we could replace Θ in each theorem with $\Theta_{\rm ach}$.

Assumption 1. *The achievable set coincides with* Θ*, that is,* $\Theta = \Theta_{\rm ach}$.

2.2 The minimax bounds: compact domains

With these administrivia out of the way, we move to delineating when distribution free minimization is possible in the case that Θ is a compact convex set. We will prove more careful results than bounds on the general minimax risk (2). For many of our lower bounds, we need only consider discrete distributions, so letting the set $\mathcal{P}_{\rm disc}(\mathcal{Z})$ be the collection of finitely supported probability distributions on $\mathcal{Z}$, define

$$\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) \coloneqq \inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}_{\rm disc}(\mathcal{Z})} \mathbb{E}_{P^n}[L_P(\widehat{\theta}_n(Z_1^n)) - L_P^\star(\Theta)].$$

We also slightly refine the risk (2) to make the results sensible. Let $\mathcal{P}_\ell(\mathcal{Z})$ be the set of Borel probability measures for which $L_P(\theta) = \mathbb{E}_P[\ell_Z(\theta)]$ is *well-defined*, meaning that

$$\begin{aligned} \mathbb{E}_P\left[|\ell_Z(\theta)|\right] < \infty \ \text{for } \theta \in \operatorname{int} \Theta \\ \mathbb{E}_P\left[[-\ell_Z(\theta)]_+\right] < \infty \ \text{ for } \theta \in \Theta. \end{aligned} \tag{9}$$

Thus we may have $L_P(\theta) = +\infty$, but $L_P(\theta) > -\infty$ and $\inf_{\theta \in \Theta} L_P(\theta) < \infty$, and L_P is convex, proper, and lower semicontinuous (Shapiro et al., 2014, Ch. 7.2.4). Then we use the refinement

$$\mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) \coloneqq \inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}_\ell(\mathcal{Z})} \mathbb{E}_{P^n}\left[L_P(\widehat{\theta}_n(Z_1^n)) - L_P^\star(\Theta)\right]$$

of the minimax risk (2). Clearly $\mathfrak{M}_n \ge \mathfrak{M}_n^{\rm low}$.

We have our first two main theorems.

Theorem 1. *Let Assumption 1 hold and Θ be compact. If Condition 1 fails, then*

$$\lim_n \mathfrak{M}_n^{\text{low}}(\ell, \mathcal{Z}, \Theta) > 0.$$

By considering estimators $\widehat{\theta}_n$ obtained by averaging n steps of the stochastic subgradient method over Θ_n, where $\Theta_n \subset \Theta$ are somewhat carefully chosen subsets (depending on $\mathcal{Z}$ and ℓ) satisfying $\Theta_n \uparrow \Theta$, we can also show a converse achievability result.

Theorem 2. *If Condition 1 holds, then*

$$\lim_n \mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) = 0.$$

This result provides no rate of convergence, which must generally depend on the structure of ℓ: without extra assumptions on ℓ, the rate may be arbitrarily slow, even in one-dimensional cases with finite $\mathcal{Z}$, as the next result shows.

Proposition 1. *Let $\mathcal{Z} = \{0, 1\}$ and $\Theta = [0, 1]$. Then there exists a numerical constant $c > 0$ such that for any continuous increasing rate function $r : [1, \infty) \to \mathbb{R}_+$, there exists a loss ℓ satisfying Condition 1 with Lipschitz constant $M([\epsilon, 1-\epsilon]) \le r^{-1}(r(1)/\epsilon)$ such that*

$$\mathfrak{M}_n^{\text{low}}(\ell, \mathcal{Z}, \Theta) \ge c\, \frac{r(1)}{r(2n)}.$$

We divide the proofs of the results into several parts, which we defer, providing an outline in Section 2.4. We provide relevant background on convexity in Section C.1 of the Appendix. We separate the arguments for Theorem 1 in the one-dimensional case that $\Theta \subset \mathbb{R}$, as the convex analytic details simplify considerably, though the intuition for the constructions remains consistent in the higher-dimensional cases; see Section C.2.1. Section C.2.2 contains the proof of Theorem 1 in the d-dimensional case. We provide the proof of achievability (Theorem 2) in Section C.3, showing its essential sharpness (Proposition 1) in Section C.4.

Before continuing, however, we present two examples to help delineate Theorems 1 and 2.

Example 3 (Log loss minimization)**:** For $z \in \mathcal{Z} = \{0, 1\}$ and $\theta \in \Theta := [0, 1]$, consider

$$\ell_z(\theta) = z \log \frac{1}{\theta} + (1 - z) \log \frac{1}{1 - \theta}.$$

The loss $\ell_z(\theta)$ is Lipschitz in θ over any compact subset interior to $[0, 1]$, though it fails to be Lipschitz at the boundaries. Nonetheless, ℓ satisfies Condition 1, so Theorem 2 shows that minimization of $\mathbb{E}[\ell_Z(\theta)]$ over $\theta \in [0, 1]$ is possible.

The log loss for estimating multinomials similarly satisfies Condition 1: for $z \in \mathcal{Z} = \{1, \ldots, k\}$ and $\Theta = \{\theta \in \mathbb{R}_+^{k-1} \mid \langle \mathbf{1}, \theta \rangle \le 1\}$, the losses

$$\ell_z(\theta) = \sum_{j=1}^{k-1} 1\{z = j\} \log \frac{1}{\theta_j} + 1\{z = k\} \log \frac{1}{1 - \langle \mathbf{1}, \theta \rangle}$$

are Lipschitz on any convex compact subsets interior to Θ. $\diamond$

Example 4 (Squared loss minimization)**:** Consider the parameter space $\Theta = [0, 1]$ and sample space $\mathcal{Z} = \mathbb{R}$, and take the loss $\ell_z(\theta) = \frac{1}{2}(\theta - z)^2$. Then $\sup_z |\ell_z'(\theta)| = \sup_z |\theta - z| = \infty$ for each θ, and Theorem 1 shows that distribution free minimization is impossible. Conversely, if $\mathcal{Z}$ is bounded then $\sup_z |\ell_z'(\theta)| < \infty$ and distribution-free minimization is possible, which highlights how the interplay between the triplet $(\ell, \mathcal{Z}, \Theta)$ determines solvability. $\diamond$

2.3 The minimax bounds: unbounded domains

Now we consider the case that Θ is unbounded, and we continue to make the standing Assumption 1 that Θ consists of achievable minimizers (8). The limiting minimax behavior in this case depends essentially on whether simultaneously for all finitely supported distributions Q, there exist compact sets attaining near minimizers. In particular, recalling $L_Q^\star(\Theta) = \inf_{\theta \in \Theta} L_Q(\theta)$, the relevant conditions become the following.

Condition 2: In addition to Condition 1, for all $\epsilon > 0$, there exists a compact $\Theta_0 \subset \Theta$ such that

$$\inf_{\theta \in \Theta_0} \left[L_Q(\theta) - L_Q^\star(\Theta)\right] \le \epsilon \tag{10}$$

for all $Q \in \mathcal{P}_{\text{disc}}(\mathcal{Z})$.

In the one-dimensional case, a simpler criterion is sufficient: that for all $\epsilon > 0$, there exists a compact $\Theta_0 \subset \Theta$ such that

$$\inf_{\theta \in \Theta_0} [\ell_z(\theta) - \ell_z^\star(\Theta)] \le \epsilon \tag{11}$$

for all $z \in \mathcal{Z}$. (See Lemma C.9 in Section C.5.) While this simplification does not appear to extend to higher dimensions, we have the following characterization.

Theorem 3. *Let Assumption 1 hold. If Condition 2 fails, then*

$$\lim_n \mathfrak{M}_n^{\text{low}}(\ell, \mathcal{Z}, \Theta) > 0.$$

We have the complementary result as well.

Proposition 2. *If Condition 2 holds, then*

$$\lim_n \mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) = 0.$$

We provide the proof of the one-dimensional case of Theorem 3 in Section C.5, demonstrating the extension to the higher dimensional case in Section C.6. Given Theorem 2, Proposition 2 admits a relatively simple proof once we demonstrate that Condition 2 immediately extends from the collection $\mathcal{P}_{\text{disc}}$ to the well-defined distributions $\mathcal{P}_\ell$ defined by the condition (9), which we provide in Section C.7. Before giving the proofs, it is instructive to consider two examples, highlighting the ways that finite Lipschitz constants are neither necessary nor sufficient for minimization to be possible.

Example 5: Consider the sample space $\mathcal{Z} = \{-1, 1\}$ and exponential loss $\ell_z(\theta) = e^{z\theta}$. While ℓ is $\mathcal{C}^\infty$, none of its derivatives are Lipschitz on $\mathbb{R}$. Nonetheless, it is clear that for any $0 < \epsilon \le 1$, we have

$$\inf_{|\theta| \le \log \frac{1}{\epsilon}} \ell_z(t) = \epsilon = \ell_z^\star + \epsilon.$$

By Proposition 2 $\mathfrak{M}_n(\ell, \mathcal{Z}, \mathbb{R}) \to 0$ as $n \to \infty$. $\diamond$

On the other hand, the absolute loss is globally Lipschitz, and yet its minimax risk over $\mathbb{R}$ is infinite:

Example 6 (Median and quantile estimation)**:** Consider the sample space $\mathcal{Z} = \mathbb{R}$ and, for a fixed $\alpha \in (0, 1)$, the α-quantile losses

$$\ell_z(\theta) := \alpha \left([\theta - z]_+ - [-z]_+\right) + (1-\alpha)\left([z - \theta]_+ - [z]_+\right),$$

where the subtraction guarantees that $L_P(\theta)$ is well-defined (9) for all distributions P on $\mathbb{R}$. Calculating left and right derivatives,

$$\begin{aligned} \mathsf{D}_+ L_P(\theta) &= P(Z \le \theta) - (1-\alpha) \\ \mathsf{D}_- L_P(\theta) &= P(Z < \theta) - (1-\alpha), \end{aligned}$$

so that if $\theta^\star$ is a $(1-\alpha)$-quantile of Z, i.e., satisfies $P(Z < \theta^\star) \le 1 - \alpha \le P(Z \le \theta^\star)$, we obtain $\mathsf{D}_- L_P(\theta^\star) \le 0 \le \mathsf{D}_+ L_P(\theta^\star)$. Nonetheless, Condition 2 fails: consider a compact interval $[t_0, t_1]$. As $\ell_z(\theta)$ is minimized at $\theta = z$, for $z \ge t_1$ we obtain

$$\inf_{\theta \le t_1} \ell_z(\theta) - \ell_z^\star = (1-\alpha)\,[z - t_1]_+,$$

which tends to $+\infty$ as $z \to \infty$, contradicting (11).

In this case, a more direct argument yields that $\mathfrak{M}_n^{\text{low}} = +\infty$, which we include only for its simplicity. Let $z_0 < z_1$ be arbitrary and to be chosen, and for $0 < \delta \le \alpha$ (also to be chosen) define the two-point distributions

$$\begin{aligned} P_0 &= (1 - \alpha + \delta)\mathbf{1}_{z_0} + (\alpha - \delta)\mathbf{1}_{z_1} \\ P_1 &= (1 - \alpha - \delta)\mathbf{1}_{z_0} + (\alpha + \delta)\mathbf{1}_{z_1}. \end{aligned}$$

Then P_0 has unique $(1-\alpha)$-quantile $\theta_0^\star = z_0$, while P_1 has unique $(1-\alpha)$-quantile $\theta_1^\star = z_1$. Recalling the bound (13), a quick calculation yields $d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) \ge \frac{\delta}{2}|z_1 - z_0|$, so

$$\mathfrak{M}_n^{\text{low}}(\ell, \mathbb{R}, \mathbb{R}) \ge \frac{\delta}{4}|z_1 - z_0|\left(1 - \|P_0^n - P_1^n\|_{\text{TV}}\right).$$

Set $\delta = \frac{1}{2n}$, so $\|P_0^n - P_1^n\|_{\text{TV}} \le n\,\|P_0 - P_1\|_{\text{TV}} \le n\delta = \frac{1}{2}$. Then taking $|z_1 - z_0| \to \infty$ gives $\mathfrak{M}_n^{\text{low}} = +\infty$. $\diamond$

In Section 3, we revisit this example in the context of finding stationary points.

2.4 Proof outlines

Each of our lower bounds follows a standard recipe of reducing optimization to a problem of testing between two distributions (Agarwal et al., 2012; Duchi, 2018; Wainwright, 2019), but because the losses have no particular form except that they cannot be uniformly Lipschitz, we must carefully construct the distributions P so that optimizing well implies testing between two statistically indistinguishable distributions. For two convex functions, define the optimization "distance"

$$d_{\mathsf{opt}}(f_0, f_1) := \inf_\theta \max\{f_0(\theta) - f_0^\star, f_1(\theta) - f_1^\star\}. \quad (12)$$

Then (Duchi, 2018, Eq. (5.2.3)) for any two (finitely supported) distributions P_0 and P_1 on $\mathcal{Z}$,

$$\mathfrak{M}_n^{\text{low}}(\ell, \mathcal{Z}, \Theta) \ge \frac{1}{2} d_{\mathsf{opt}}(L_{P_0}, L_{P_1})(1 - \|P_0^n - P_1^n\|_{\text{TV}}) \quad (13)$$

Our lower bound proofs thus pursue competing goals: to construct well-separated losses L_{P_0} and L_{P_1} while the distributions P_0, P_1 generating them are too close to distinguish. We give an overview of the constructions here, leaving details to the proofs in Section C.

Let us focus for now on the case that Condition 1 fails, so that there are points $\theta_0 \in \operatorname{int}\Theta$ with arbitrarily large gradient magnitude $M_{\theta_0} := \sup_{z \in \mathcal{Z}} \sup_{g \in \partial \ell_z(\theta_0)} \|g\|_2$. Let $z_0 \in \mathcal{Z}$ and $g \in \partial \ell_{z_0}(\theta_0)$ be one of these arbitrarily large gradients, defining the normalized version $v = g / \|g\|_2$. Then because of the standing Assumption 1, taking $t = \langle v, \theta_0 \rangle + \alpha$ for some constant $\alpha > 0$, the half-space $\{\theta \mid \langle v, \theta \rangle \ge t\}$ contains what we term a *directable* set C (e.g., a finitely supported distribution Q defining a compact set $\theta^\star(Q)$ of minimizers, as in Definition 2.2 and Lemma A.3; see Fig. 3). Take this point θ_0 and the direction v, and let Q be the finitely supported distribution so that L_Q satisfies the inclusion in Definition 2.2; as Figure 1 illustrates, we have effectively "placed" a desired gradient at some distance from θ_0.

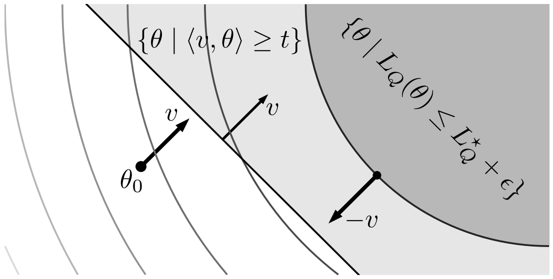

Figure 1. Constructing the lower bound. At the point θ_0, $\ell_z(\theta_0)$ has arbitrarily large gradient g in the direction v. The sublevel set $L_Q(\theta) \leq L_Q^\star + \epsilon$, contained in the halfspace $\{\theta \mid \langle v, \theta\rangle \geq t\}$ strictly separated from θ_0, has a point with gradient $\nabla L_Q(\theta) = -v$.

To optimize L_Q to accuracy better than some constant c, one must produce θ satisfying $\langle v, \theta\rangle \geq t$, as Figure 1 evidences. On the other hand, defining the mixture distribution

$$P_1 \coloneqq \left(1 - \frac{1}{n}\right) Q + \frac{1}{n} \mathbf{1}_{z_0}$$

that "hides" the observation z_0 by giving it small probability, to optimize the corresponding loss $L_{P_1}(\theta) = (1 - \frac{1}{n}) L_Q(\theta) + \frac{1}{n} \ell_{z_0}(\theta)$ to accuracy better than a constant, we must have $\langle v, \theta\rangle < t$, as otherwise, the loss $\ell_{z_0}(\theta)$ must grow arbitrarily. More precisely, at any θ with $\langle v, \theta\rangle \geq t$ we have $\ell_{z_0}(\theta) \geq \ell_{z_0}(\theta_0) + \|g\|_2 \langle v, \theta - \theta_0\rangle$. The rightmost term is now larger than $\|g\|_2 \alpha$ by construction, so we can make it arbitrarily large by choosing $\|g\|_2$ to be large. In particular, we see that

$$d_{\mathsf{opt}}(L_Q, L_{P_1}) \geq c$$

for a constant $c > 0$. Now that we have $\|Q - P_1\|_{\mathrm{TV}} \leq \frac{1}{n}$ we need an upper bound on $\|Q^n - P_1^n\|_{\mathrm{TV}}$ to apply inequality (13). The next lemma (whose standard proof we provide for completeness in Appendix B.2) allows precise control over such mixtures.

Lemma 2.1. *Let P_0 and P_1 be distributions satisfying $\|P_0 - P_1\|_{\mathrm{TV}} \leq \gamma$. Then*

$$\|P_0^n - P_1^n\|_{\mathrm{TV}} \leq \sqrt{1 - (1 - \gamma)^{2n}}.$$

In particular, if $P_{0,n}$ and $P_{1,n}$ are sequences of distributions with $\|P_{0,n} - P_{1,n}\|_{\mathrm{TV}} \leq a/n$, where $a < \infty$, then

$$\limsup_n \left\|P_{0,n}^n - P_{1,n}^n\right\|_{\mathrm{TV}} \leq \sqrt{1 - e^{-2a}}.$$

Because $\|Q^n - P_1^n\|_{\mathrm{TV}} \leq \sqrt{1 - (1 - 1/n)^{2n}} \to \sqrt{1 - e^{-2}}$, inequality (13) then implies

$$\mathfrak{M}_n^{\mathrm{low}}(\ell, \mathcal{Z}, \Theta) \geq \frac{c}{2}\left(1 - \sqrt{1 - e^{-2}}\right).$$

Section C.2 makes each of these steps rigorous for the compact case. In the case where Θ is unbounded, we require a bit more care to generate the separated losses, which we obtain not by making an arbitrarily large gradient but instead by placing the directable set arbitrarily far away from any given θ_0. See Section C.5 for the details.

3 ATTAINING STATIONARY POINTS

In many M-estimation problems, instead of minimizing the loss L_P, we instead seek (near) stationary points of L_P over $\Theta = \mathbb{R}$. This has many motivations, including in non-convex problems Nesterov (2012); Lee et al. (2016); Arjevani et al. (2023); Bubeck and Mikulincer (2020), where attaining strong minimization guarantees is impossible, as well as recent work that connects "gradient equilibrium" to various desiderata in prediction problems Angelopoulos et al. (2025), including online sequence calibration Foster and Vohra (1998); Foster and Hart (2021) and conformal prediction Gibbs and Candès (2021); Angelopoulos et al. (2023). The simplest motivating example comes by revisiting Example 6 on quantile estimation, where Theorem 3 shows that minimizing the loss without any assumptions is impossible.

Example 7 (Quantile estimation, Example 6 revisited)**:** For the losses $\ell_\alpha(\theta, z) = \alpha\,[\theta - z]_+ + (1 - \alpha)\,[z - \theta]_+$, we see that if

$$-\epsilon \leq \mathsf{D}_- L_P(\theta) \quad \text{and} \quad \mathsf{D}_+ L_P(\theta) \leq \epsilon,$$

then evidently

$$1 - \alpha - \epsilon \leq P(Z < \theta) \quad \text{and} \quad P(Z \leq \theta) \leq 1 - \alpha + \epsilon,$$

so that finding a nearly stationary point—one with small left and right derivatives—implies an accurate quantile estimate. As we shall see in the sequel, any empirical minimizer $\widehat{\theta}_n = \mathrm{argmin}_\theta\, L_{P_n}(\theta)$ (i.e., empirical quantile) satisfies $\max\{\mathsf{D}_+ L_P(\widehat{\theta}_n), -\mathsf{D}_- L_P(\widehat{\theta}_n)\} \leq O(1)/\sqrt{n}$ with high probability, so that

$$P(Z_{n+1} \leq \widehat{\theta}_n) \geq 1 - \alpha - \tfrac{O(1)}{\sqrt{n}}$$
$$P(Z_{n+1} < \widehat{\theta}_n) \leq 1 - \alpha + \tfrac{O(1)}{\sqrt{n}}.$$

where Z_{n+1} is an independent draw from P. This contrasts strongly with the impossibility result for loss *minimization* in Example 6. ◇

In this section, we expand this example to consider 1-dimensional Lipschitz convex losses. We view these results as essentially preliminaries and (hopefully) starting points for future work. We will not provide fundamental lower bounds, instead treating these results as evidence (see also Angelopoulos et al.'s discussion (Angelopoulos et al., 2025)) that there are significant differences between finding (near) stationary points and even asymptotically minimizing convex losses.

3.1 The definition of stationarity

When the space Θ is unbounded, Condition 2 and Theorem 3 show that "typically" $\lim_n \mathfrak{M}_n^{\mathrm{low}}(\ell, \mathcal{Z}, \Theta) > 0$, meaning that distribution free estimation is impossible. Accordingly, we modify the goals to seek *stationary points* of L_P. If L were differentiable and had minimizers, a natural metric for stationarity would be the derivative magnitude

$$\epsilon_L(\theta) := |L'(\theta)|.$$

This is not quite satisfactory, because (as in the quantile example 7) we frequently seek stationary points of non-differentiable losses, and additionally, it is possible that L has no (finite) minimizer. Accordingly, we give an expanded definition for error, which coincides when L is differentiable and has attained minimizers, but allows moving beyond these cases. As a first attempt to address non-differentiability, we might mimic Example 7 to define

$$\epsilon_L(\theta) := \left[\max\left\{-\mathsf{D}_+ L(\theta), \mathsf{D}_- L(\theta)\right\}\right]_+ .$$

For example, when $L(\theta) = |\theta|$, then $\theta_0 = \theta_1 = 0$, and $\mathsf{D}_+ L(0) = 1$ while $\mathsf{D}_- L(0) = -1$, and $\epsilon_L(\theta) = 0$ if and only if $\theta = 0$; more generally, if $\theta^\star$ minimizes L, then we know that $\mathsf{D}_- L(\theta^\star) \le 0 \le \mathsf{D}_+ L(\theta^\star)$ and $\epsilon_L(\theta^\star) = 0$. When minimizers of L may not exist, we incorporate a correction subtracting off the minimal possible subgradient magnitude, defining

$$\epsilon_L(\theta) := [\max\{-\mathsf{D}_+ L(\theta), \mathsf{D}_- L(\theta)\}]_+ - \inf_\theta \inf_{g \in \partial L(\theta)} \{|g|\}. \tag{14}$$

In the case that L is differentiable and has a minimizer, we simply have $\epsilon_L(\theta) = |L'(\theta)|$. Figure 2 shows the help of the correction term (14).

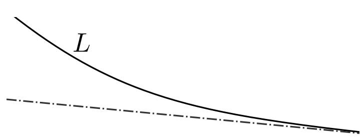

Figure 2. The function L decreases out to ∞, but points to the right are nearly stationary, approaching the asymptotically shallowest slope $\inf_\theta |L'(\theta)|$.

A few properties of ϵ_L show that it provides a natural measure of non-stationarity.

Lemma 3.1. *Let $L : \mathbb{R} \to \overline{\mathbb{R}}$ be closed convex. The stationarity error* (14) *satisfies*

(i) If $0 \in \partial L(\theta)$, then $\epsilon_L(\theta) = 0$.

(ii) If $0 \notin \partial L(\theta)$ and argmin L *is non-empty, then $\epsilon_L(\theta) > 0$.*

(iii) If argmin L *is empty, then either $\mathsf{D}_+ L(\theta) < 0$ and $\mathsf{D}_- L(\theta) < 0$ for all θ or $\mathsf{D}_- L(\theta) > 0$ and $\mathsf{D}_+ L(\theta) > 0$ for all θ. In either case, the choice $s = -\operatorname{sign}(\mathsf{D}_- L(\theta))$ yields*

$$\lim_{\theta \to s \cdot \infty} \epsilon_L(\theta) = 0.$$

See Appendix D.1 for a proof.

3.2 Achieving stationarity

When the loss ℓ is always (globally) Lipschitz, empirical risk minimizers are sufficient to achieve stationarity according to the measure (14), even when losses are non-differentiable (see Appendix D.2 for a proof).

Proposition 3. *Assume that $\ell_z(\cdot)$ is M-Lipschitz for all $z \in \mathcal{Z}$. Then any empirical minimizer $\widehat{\theta} \in \operatorname{argmin}_\theta L_{P_n}(\theta)$ satisfies*

$$\mathbb{P}\left(\epsilon_{L_P}(\widehat{\theta}) > t\right) \le 2 \exp\left(-\frac{nt^2}{2M^2}\right) \quad \text{for } t \ge 0.$$

As an immediate consequence of Proposition 3

$$\mathbb{E}\left[\epsilon_P(\widehat{\theta}_n)\right] \le 2 \int_0^\infty \exp\left(-\frac{nt^2}{2M^2}\right) dt = \frac{2\sqrt{\pi} M}{\sqrt{n}}.$$

Let us consider this in the context of Example 7: for the quantile loss,

$$\epsilon_{L_P}(\theta) = \max\left\{(1-\alpha) - P(Z \le \theta), P(Z < \theta) - (1-\alpha)\right\}.$$

So with probability at least $1 - 2\exp(-t^2/2)$,

$$P(Z_{n+1} \le \widehat{\theta}_n) \ge 1 - \alpha - \tfrac{t}{\sqrt{n}},$$
$$P(Z_{n+1} < \widehat{\theta}_n) \le 1 - \alpha + \tfrac{t}{\sqrt{n}}.$$

In this particular case, the quantile satisfies a slightly sharper guarantee (Duchi, 2025, Proposition 2), but Proposition 3 holds for general Lipschitz functions.

4 DISCUSSION

This paper has developed the conditions that provide a concrete line between those cases in which distribution-free minimization of convex losses is asymptotically possible and when, conversely, it is not. Here, we situate some of the results relative to the statistical and machine learning literatures while enumerating several open questions.

In classical (stochastic) optimization and machine learning, uniform convergence guarantees (Devroye et al., 1996; Wainwright, 2019; Bartlett and Mendelson, 2002) of the form $\sup_{\theta\in\Theta} |L_{P_n}(\theta) - L_P(\theta)| \to 0$ (in probability or otherwise) provide a frequent tool for demonstrating the convergence of the empirical risk minimizer $\widehat{\theta}_n = \mathop{\rm argmin}_{\theta\in\Theta} L_{P_n}(\theta)$. As we mention in the introduction, they are sometimes equivalent to the convergence $\mathfrak{M}_n \to 0$ of the minimax risk (Alon et al., 1997; Vapnik, 1998; Anthony and Bartlet, 1999; Shalev-Shwartz et al., 2010).

Condition 1 shows that even when Θ is compact, these are unnecessary for estimation, and we can even have $\sup_{\theta\in\Theta} |L_{P_n}(\theta) - L_P(\theta)| = \infty$ with probability 1 while $L_P(\widehat{\theta}_n) \to L_P^\star$. Indeed, in Example 3, let $p = P(Z = 1)$. Then whenever $\widehat{p} = \frac{1}{n}\sum_{i=1}^n 1\{Z_i = 1\} \neq p$,

$$\sup_{\theta\in[0,1]} |L_{P_n}(\theta) - L_P(\theta)| = \sup_{\theta\in[0,1]} \left|(\widehat{p}_n - p)\log\tfrac{\theta}{1-\theta}\right| = \infty.$$

If p is irrational this occurs at all n, while Theorem 2 shows that there exists $\widehat{\theta}_n$ achieving $L_P(\widehat{\theta}_n) - L_P^\star = o_P(1)$. In the unbounded case, Example 5 makes clear that no form of global Lipschitz continuity (and similarly, no uniform convergence) is necessary for convergence of an estimator. Nonetheless, it would be interesting to understand whether the existence of stable estimators (say, that change little upon substituting a single observation Z_i) is necessary and sufficient for ℓ to be minimizable, as is the case for bounded functions in machine learning problems (Shalev-Shwartz et al., 2010). It also remains unclear whether Condition 1 is equivalent to the consistency of particular types of regularization.

Section 3 implicitly raises the question of whether our choice of minimax risk (2) via the excess loss $L_P(\theta) - L_P^\star$ is the right one. In some problems, stationary points may be more interesting (and achievable) than loss minimizers. In another vein, our characterizations are loss specific but do not address questions of adapting to the underlying distribution P (cf. Chatterjee et al. (2016)): are instance-specific guarantees, i.e., depending on ℓ, P, and Θ, possible? Relatedly, if instead of the minimax risk (2) we consider a game where we choose an estimator $\widehat{\theta}_n$ for each n, then nature chooses P and we observe $Z_i \stackrel{\rm iid}{\sim} P$, we ask whether for some $\gamma \geq 0$ the risk measure

$$\inf_{\widehat{\theta}} \sup_P \limsup_n n^\gamma \cdot \mathbb{E}_{P^n}\left[L_P(\widehat{\theta}_n(Z_1^n)) - L_P^\star\right] \qquad (15)$$

converges. Super-efficiency theory (van der Vaart, 1997) shows that estimators may improve upon classical statistical efficiency bounds only at a measure-zero set of parameters to provide insights into the limit (15) in these cases; extending it to general loss minimization problems would be a major achievement.

We could instead ask for convergence of $\widehat{\theta}_n$ to the solution set $\Theta^\star(P) := \mathop{\rm argmin}_{\theta\in\Theta} L_P(\theta)$. When Θ is compact, the empirical risk minimizer $\widehat{\theta}_n$ satisfies $\mathop{\rm dist}(\widehat{\theta}_n, \Theta^\star(P)) \stackrel{a.s.}{\to} 0$ for any P, but this convergence is pointwise (Shapiro et al., 2014, Thm. 5.4). Even for Lipschitz functions on a compact set, uniform guarantees are impossible: consider minimizing $\ell_z(\theta) = |\theta - z|$, and take the two distributions $P_0 = \frac{1+\delta}{2}\mathbf{1}_0 + \frac{1-\delta}{2}\mathbf{1}_1$ and $P_1 = \frac{1-\delta}{2}\mathbf{1}_0 + \frac{1+\delta}{2}\mathbf{1}_1$. Then because $\theta_0 = \mathop{\rm argmin} L_{P_0} = 0$ and $\theta_1 = \mathop{\rm argmin} L_{P_1} = 1$, an essentially standard minimax calculation (Wainwright, 2019, Ch. 14) gives

$$\inf_{\widehat{\theta}_n} \max_{i\in\{0,1\}} \mathbb{E}_{P_i^n}\left[|\widehat{\theta}_n(Z_1^n) - \theta_i|\right] \geq \tfrac{1}{2}\left(1 - \sqrt{\delta^2(1 + o_\delta(1))\, n}\right).$$

where the final inequality uses Pinsker's inequality and $D_{\rm kl}(P_0^n \| P_1^n) = n D_{\rm kl}(P_0 \| P_1)$ while $D_{\rm kl}(P_0 \| P_1) = \delta \log\frac{1+\delta}{1-\delta} = 2\delta^2(1 + o(1))$ as $\delta \to 0$. Take $\delta \ll 1/n$ to obtain that the minimax parameter estimation risk is as bad as predicting $\widehat{\theta} = 0$: $\max_{i\in\{0,1\}} \mathbb{E}_{P_i}[|\widehat{\theta}_n - \theta_i|] \geq \frac{1}{2} - o(1)$. Any characterization of estimable parameters in a distribution free sense requires care, as certainly Condition 1 is insufficient.

We identify one more important open question. First, we include Section 3 only to contrast the (in)achievability results present in the rest of the paper. Providing a characterization of the losses for which it is possible to attain stationary points in a distribution free sense remains a challenging. Certainly, uniform Lipschitz continuity of the losses (as Proposition 3 assumes) is unnecessary: if $\ell_z(\cdot)$ has Lipschitz derivatives on compact subsets of Θ, then any estimator with $L_P(\widehat{\theta}_n) - L_P^\star \to 0$ guarantees that $L_P'(\widehat{\theta}_n) \to 0$. Indeed, if $L_P(\theta) - L_P^\star \leq \epsilon$, then because a function with M-Lipschitz derivative satisfies $|L_P(\theta + t) - L_P(\theta) - L_P'(\theta)t| \leq Mt^2/2$, taking $t = -L_P'(\theta)/M$ yields

$$L_P(\theta + t) \leq L_P(\theta) - \frac{L'(\theta)^2}{2M}.$$

That is, necessarily $L'(\theta) \leq \sqrt{2M\epsilon}$. So, for example, the exponential loss in Example 5 admits estimators satisfying $L_P'(\widehat{\theta}_n) \to 0$, though it is certainly not Lipschitz. We leave such extensions to future work.

Acknowledgements

Partially supported by the Office of Naval Research Grant N00014-22-1-2669. This work began as part of the "Algorithmic Stability" workshop hosted by the American Institute of Mathematics (AIM), May 2025.

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the Association for Computing Machinery*, 44(4):615–631, 1997.

A. Angelopoulos, E. Candès, and R. J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems 36*, 2023.

A. N. Angelopoulos, M. I. Jordan, and R. J. Tibshirani. Gradient equilibrium in online learning: Theory and applications. *arXiv:2501.08330 [cs.LG]*, 2025.

M. Anthony and P. Bartlet. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming, Series A*, 199:165–214, 2023.

R. Bahadur and L. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10 (2):455–482, 2021.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, California, 1986.

S. Bubeck and D. Mikulincer. How to trap a gradient flow. In *Proceedings of the Thirty Third Annual Conference on Computational Learning Theory*, pages 940–960, 2020.

S. Chatterjee, J. Duchi, J. Lafferty, and Y. Zhu. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. Donoho. One-sided inference about functionals of a density. *Annals of Statistics*, 16(4):1390–1420, 1988.

J. C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

J. C. Duchi. A few observations on sample-conditional coverage in conformal prediction. *arXiv:2503.00220 [math.ST]*, 2025.

D. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

D. P. Foster and S. Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12), 2021.

I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems 34*, 2021.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993a.

J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II*. Springer, New York, 1993b.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Proceedings of the Twenty Ninth Annual Conference on Computational Learning Theory*, 2016.

J. Lei. Classification with confidence. *Biometrika*, 101 (4):755–769, 2014.

J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 76(1):71–96, 2014.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. How to make the gradients small. *Optima*, 88, 2012.

M. Raginsky and A. Rakhlin. Information-based complexity, feedback, and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11: 2635–2670, 2010.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Optimization Society, second edition, 2014.

C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–620, 1977.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A. W. van der Vaart. Superefficiency. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, chapter 27. Springer, 1997.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

V. Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2):349–376, 2013. doi: 10.1007/s10994-013-5355-6. URL https://doi.org/10.1007/s10994-013-5355-6.

V. Vovk, A. Grammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Checklist

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

Distribution Free M-estimation: Supplementary Materials

A PROPERTIES OF THE SET OF ACHIEVABLE MINIMIZERS

With the definitions of directional derivatives and associated calculus rules in Section C.1, we can provide more perspective on the achievable set (8).

A.1 The one-dimensional achievable set

The cleanest results follow in the one-dimensional case, as the achievable set (6) admits many different characterizations. To do so, we first argue that so long as Θ has an interior, the extreme points (5) satisfy

$$\theta_{\min} = \inf\left\{\theta \in \Theta \mid \sup_{z \in \mathcal{Z}} \mathsf{D}_+\ell_z(\theta) > 0\right\} \quad \text{and} \quad \theta_{\max} = \sup\left\{\theta \in \Theta \mid \inf_{z \in \mathcal{Z}} \mathsf{D}_-\ell_z(\theta) < 0\right\}. \tag{16}$$

We show the left equality, as the right is similar. Let $\theta \in \operatorname{int}\Theta$ be any point for which $\sup_{z\in\mathcal{Z}} \mathsf{D}_+\ell_z(\theta) > 0$. Then there exists $z \in \mathcal{Z}$ for which $\mathsf{D}_+\ell_z(\theta) > 0$, and because $\ell_z(\cdot)$ is a.e. differentiable on its domain, for all $\epsilon > 0$ there exists $\theta' \in [\theta, \theta + \epsilon]$ for which $\ell_z'(\theta') \ge \mathsf{D}_+\ell_z(\theta)$ by Lemmas C.1 and C.2, so the claimed equality holds if there exists $\theta \in \operatorname{int}\Theta$ satisfying $\mathsf{D}_+\ell_z(\theta) > 0$ for some $z \in \mathcal{Z}$. If, on the other hand, $\sup_{z\in\mathcal{Z}} \mathsf{D}_+\ell_z(\theta) \le 0$ for all $\theta \in \operatorname{int}\Theta$, then the monotonicity of D_+ and ℓ' show that $\theta_{\min} = \sup\{\theta \in \Theta\}$ as desired. We therefore use the characterization (16) for $\theta_{\min}$ and $\theta_{\max}$.

While the definition (8) of Θ_{ach} appears complex, the next two lemmas show how it captures the sets of plausible minimizers of $L_P(\theta)$ in ways that depend only on the loss ℓ and potential data $\mathcal{Z}$, irrespective of P. In the lemmas, we recall that $\mathcal{P}_\ell$ denotes the set (9) of Borel probabilities for which L_P is well defined on $\theta \in \Theta$. All finitely supported measures P immediately belong to $\mathcal{P}_\ell$, and moreover, minimizers of L_P in Θ are attained. If Θ is compact, this is standard. Otherwise, the extended reals address the issue, so that (e.g.) if $\sup\Theta = +\infty$ and $\lim_{\theta\to\infty} L(\theta) = \inf_{\theta\in\Theta} L(\theta)$, we say $\infty \in \operatorname{argmin}_{\theta\in\Theta} L(\theta)$.

Lemma A.1. *Simultaneously for each probability distribution $P \in \mathcal{P}_\ell$, that is, for which L_P is well-defined on Θ, the following hold:*

(i) If $\theta_{\min} \ge \theta_{\max}$, then $\theta_{\max} \le \theta^\star \le \theta_{\min}$ implies that $\theta^\star$ minimizes $L_P(\theta)$ over $\theta \in \Theta$. In particular, all functions $\ell_z(\cdot)$ are constant and minimized on $[\theta_{\max}, \theta_{\min}]$.

If Θ_{ach} is non-empty, we additionally have

(ii) Θ_{ach} intersects the minimizers of $L_P(\theta)$, that is, $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_P[\ell_Z(\theta)] \cap \Theta_{\mathrm{ach}} \neq \emptyset$.

(iii) For any point $\theta_0 \in \Theta$, the projection $\mathsf{proj}_{\Theta_{\mathrm{ach}}}(\theta_0) = \operatorname{argmin}_{\theta\in\Theta_{\mathrm{ach}}}\{(\theta - \theta_0)^2\}$ satisfies $L_P(\mathsf{proj}_{\Theta_{\mathrm{ach}}}(\theta_0)) \le L_P(\theta_0)$.

(iv) If Θ_{ach} is non-singleton, then

$$\Theta_{\mathrm{ach}} = \operatorname{cl}\left\{\theta \in \Theta \mid \text{ there exist } z_0, z_1 \in \mathcal{Z} \text{ with } \begin{array}{l} \partial\ell_{z_0}(\theta) \cap \mathbb{R}_{>0} \neq \emptyset \quad \text{and} \\ \partial\ell_{z_1}(\theta) \cap \mathbb{R}_{<0} \neq \emptyset \end{array}\right\}. \tag{17}$$

We prove the lemma in Appendix B.5.

We can write Θ_{ach} more directly in terms of minimizers of L_P. Let $\mathcal{A}(\ell, \mathcal{Z}, \Theta)$ be the collection of closed convex sets $C \subset [-\infty, \infty]$ for which for all distributions P on $\mathcal{Z}$,

$$C \cap \operatorname*{argmin}_{\theta\in\Theta} L_P(\theta) \neq \emptyset.$$

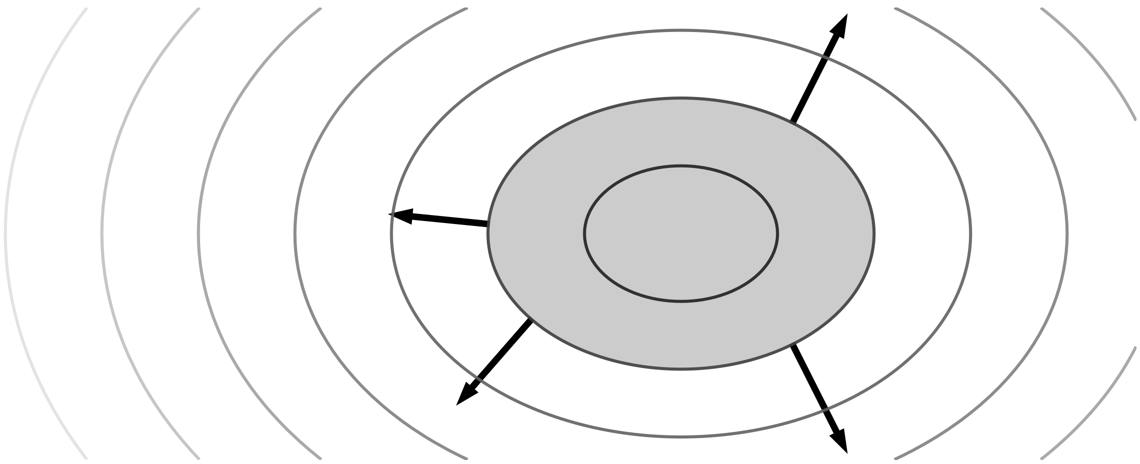

Figure 3. A directable set constructed as the sublevel set of the function $\ell(\theta) = \frac{1}{2}\theta_1^2 + \theta_2^2$. The gradients point out of the sublevel set $\{\theta \mid \frac{1}{2}\theta_1^2 + \theta_2^2 \le c\}$ and have norm $\|\nabla \ell(\theta)\|_2$ bounded away from 0.

We then have the following result, whose proof we provide in Appendix B.6, and which makes formal that the definitions (6) and (8) are equivalent.

Lemma A.2. *Assume Θ_{ach} is non-empty. Then viewed as a subset of the extended reals $[-\infty, \infty]$, it satisfies*

$$\Theta_{\text{ach}} = \bigcap \{C \in \mathcal{A}(\ell, \mathcal{Z}, \Theta)\}.$$

That is, Θ_{ach} is the smallest closed convex set C satisfying $C \cap \operatorname{argmin}_{\theta\in\Theta} L_P(\theta) \neq \emptyset$ for all P. In summary, Lemmas A.1 and A.2 show that it is no loss of generality to restrict attention to the set (8) of achievable minimizers: completely independently of the distribution P on $\mathcal{Z}$, we can *always* improve the expected loss L_P by projecting a point onto Θ_{ach}.

A.2 Characterizations of directable sets

In this section we give a few sufficient characterizations for directability to show that such sets are a typical phenomenon. The first condition is that any time a loss attains its minimizers on a compact set, the set of minimizers is directable. See Figure 3 for an illustration.

Lemma A.3. *Let Q be finitely supported and assume $\theta^\star(Q) := \operatorname{argmin}_{\theta\in\Theta} L_Q(\theta)$ is compact. Then $\theta^\star(Q)$ is directable.*

This lemma is an immediate consequence of Proposition 4 to come in Section C.1.3.

We can also more directly analogize the condition (5). Recall the *directional derivative* (Hiriart-Urruty and Lemaréchal, 1993a, Def. VI.1.1.1) of a convex function f at the point $\theta \in \operatorname{dom} f$ in the direction v,

$$\mathsf{D}f(\theta; v) := \lim_{t\downarrow 0} \frac{f(\theta + tv) - f(\theta)}{t} = \inf_{t>0} \frac{f(\theta + tv) - f(\theta)}{t} \stackrel{(\star)}{=} \sup_{g\in\partial f(\theta)} \langle g, v\rangle$$

where the limit always exists, $\mathsf{D}f(\theta; v)$ is convex and positively homogeneous in v, and the equality $(\star)$ holds if $\partial f(\theta) \neq \emptyset$. If f is differentiable at θ, then $\mathsf{D}f(\theta; v) = \langle \nabla f(\theta), v\rangle$. We recall our Definition 2.1 that a point θ is *achievable* if for each $v \in \mathbb{S}^{d-1}$,

$$\text{there exists } z \in \mathcal{Z} \ \text{ with } \ \mathsf{D}\ell_z(\theta; v) > 0.$$

The next lemma shows that achievable points are also directable (see Appendix B.1 for proof).

Lemma A.4. *If $\theta \in \operatorname{int}\Theta$ is achievable (Definition 2.1), then $\{\theta\}$ is directable (Definition 2.2). In the Definition 2.2, we may take the set $C_\epsilon = \{\theta\}$ for all $\epsilon > 0$.*

A.3 Characterizations of the achievable set in $d > 1$ dimensions

When $\Theta \subset \mathbb{R}^d$ for $d > 1$, the characterizations of the achievable set (8) are not quite so elegant as those for the one-dimensional case, so instead we simply record a few results and extensions to the set. As a trivial remark, note that if $\theta_0 \notin \Theta_{\text{ach}}$, then for each finitely supported distribution Q, there is $\theta \in \Theta_{\text{ach}}$ for which $L_Q(\theta) \le L_Q(\theta_0)$, so in that sense, there is no reason to include points in $\Theta \setminus \Theta_{\text{ach}}$ in estimation.

As a preliminary result, we note that the condition that the halfspaces are unconstraining (7) extends to any probability measure so long as the loss ℓ_z is appropriately integrable.

Lemma A.5. *Let $\mathcal{P}_\ell$ contain the well-defined probability measures (9) and Θ be compact. Then equality (7) holds for all $P \in \mathcal{P}_\ell$.*

See Appendix B.7 for a proof. In brief, we see that the restriction to finitely supported measures is of no real import.

We also provide a few equivalent characterizations of achievability, assuming the density of achievable points (Definition 2.1). For a vector θ_0, define the set of *eliminatable points*

$$E(\theta_0, \ell) \coloneqq \{\theta \mid \mathsf{D}\ell_z(\theta_0; \theta - \theta_0) \ge 0 \text{ for all } z \in \mathcal{Z}\} = \bigcap_{z \in \mathcal{Z}} \{\theta \mid \mathsf{D}\ell_z(\theta_0; \theta - \theta_0) \ge 0\}, \tag{18}$$

which is a convex cone of points that necessarily have loss at least that of θ_0, as

$$\ell_z(\theta) \ge \ell_z(\theta_0) + \mathsf{D}\ell_z(\theta_0; \theta - \theta_0).$$

If a point $\theta \in E(\theta_0, \ell)$ for some $\theta_0 \ne \theta$, then *a priori* θ is sub-optimal (relative to θ_0) for any probability distribution P on $\mathcal{Z}$, and $L_P(\theta_0) \le L_P(\theta)$. In fact, that a point θ_0 eliminates no points in Θ is equivalent to its achievability (see Appendix B.8):

Lemma A.6. *If $\Theta \cap E(\theta_0, \ell) = \{\theta_0\}$ for some $\theta_0 \in \operatorname{int}\Theta$, then Definition 2.1 holds at θ_0. Conversely, if the inequality in Definition 2.1 holds at $\theta \in \operatorname{int}\Theta$, then $\theta \notin E(\theta_0, \ell)$ for all $\theta_0 \in \operatorname{int}\Theta$ with $\theta_0 \ne \theta$.*

As a consequence, if each $\theta \in \operatorname{int}\Theta$ is achievable (Definition 2.1), then no points may be summarily eliminated as sub-optimal from Θ, and $\Theta_{\text{ach}} = \Theta$.

An alternative to the characterizations above is to define the set

$$\mathsf{domin}_\ell(\theta_0) \coloneqq \{\theta \in \Theta \mid \ell_z(\theta) \le \ell_z(\theta_0) \text{ for all } z \in \mathcal{Z}\} = \bigcap_{z \in \mathcal{Z}} \{\theta \in \Theta \mid \ell_z(\theta) \le \ell_z(\theta_0)\}$$

of points that uniformly dominate θ_0, meaning that $\theta \in \mathsf{domin}_\ell(\theta_0)$ achieves smaller loss on *all* examples $z \in \mathcal{Z}$. The set $\mathsf{domin}_\ell(\theta_0)$ is closed convex, as it is the intersection of closed convex sets. When the only point θ that uniformly dominates θ_0 is always θ_0 itself, then it turns out that all points in the interior of Θ are achievable (Definition 2.1):

Lemma A.7. *Assume that $\mathsf{domin}_\ell(\theta_0) = \{\theta_0\}$ for each $\theta_0 \in \operatorname{int}\Theta$. Then for each $\theta \in \operatorname{int}\Theta$ and $v \in \mathbb{S}^{d-1}$, there exists $z \in \mathcal{Z}$ such that*

$$\mathsf{D}\ell_z(\theta; v) \ge \mathsf{D}_-\ell_z(\theta; v) > 0.$$

Conversely, if $\theta_0 \in \operatorname{int}\Theta$ is achievable (Definition 2.1), then $\mathsf{domin}_\ell(\theta_0) = \{\theta_0\}$.

See Appendix B.9 for a proof.

Recalling Example 1, we see what we view as the prototypical case: the set of achievable points (Definition 2.1) is dense in $\operatorname{int}\Theta$. When this occurs, the somewhat complicated seeming achievable set (8) coincides with the closure of any of the following sets:

(i) the admissible points $\theta \in \Theta$, i.e., those belonging to the set defined in (8).

(ii) the points θ that are not eliminatable (18), that is, $\theta \notin E(\theta_0, \ell)$ for any $\theta_0 \in \operatorname{int}\Theta$.

(iii) the set of un-dominated points $\theta \in \Theta$, i.e., those for which $\mathsf{domin}_\ell(\theta) = \{\theta\}$.

A.4 Further comments on solvability without Assumption 1

In this section we highlight the fact that in the pathological cases where $\Theta \neq \Theta_{\text{ach}}$, then Conditions 1 and 2 may not correctly determine if an instance is solvable. In fact, we can construct counterexamples where our conditions fail to be the correct criteria.

Example 8: Let $\mathcal{Q}$ be the set of rational numbers, take $\mathcal{Z} = [0,1] \setminus \mathcal{Q} \times \mathcal{Q}$, and $\Theta = [0,2]$. Consider the losses indexed by $z = (r,q)$ via

$$\ell_{r,q}(\theta) := \mathbf{1}\{\theta \leq 1\} \frac{1}{|r-q|} \cdot \frac{1}{\theta} + \mathbf{1}\{\theta > 1\} \left(\theta - 1 + \frac{1}{|r-q|}\right).$$

For this triplet $(\mathcal{Z}, \Theta, \ell)$ Condition 1 fails, as the smallest Lipschitz constant at each $\theta < 1$ is $+\infty$, but $\theta = 1$ trivially minimizes each loss ℓ_z. $\diamond$

We can also recreate the same phenomenon beyond one dimension.

Example 9: Once again take $\mathcal{Z} = [0,1] \setminus \mathcal{Q} \times \mathcal{Q}$, and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$, with losses

$$\ell_{r,q}(\theta) = \frac{1}{|r-q|} \|\theta\|_2 .$$

Evidently, Condition 1 fails once again, but $\theta = 0$ trivially minimizes each loss ℓ_z. In this case, every directable set (Definition 2.2) contains $\{0\}$, so the achievable set (8) is $\{0\}$. $\diamond$

As we discuss in the second half of Section 2.1, for general $d > 1$, achievable minimizers (Definition 2.1) provide the cleanest description of solvable problems. The more elaborate Definition 2.2 addresses issues in higher dimensions that arise because there is no natural ordering in $\mathbb{R}^d$, so gradients are only monotonic in one direction at a time. Lemma A.4 indicates that the achievable set (8) includes the convex hull of these achievable points. Even further, in the case of GLMs or the case of location-type losses, it is the convex hull of achievable points.

B TECHNICAL PROOFS

B.1 Proof of Lemma A.4

We first perform a compactness argument when $\theta \in \operatorname{int} \Theta$ that refines the positivity condition in Definition 2.1. Because $\partial \ell_z(\theta)$ is compact for $\theta \in \operatorname{int} \Theta$, $v \mapsto \mathsf{D}\ell_z(\theta; v)$ is convex and bounded on compacts, and so is continuous. Thus, for each v, for the z satisfying Definition 2.1, there is $\epsilon = \epsilon(v) > 0$ such that

$$\inf_{\|u-v\|_2 < \epsilon, \|u\|_2 = 1} \mathsf{D}\ell_z(\theta; u) > 0.$$

The sets $\{v + \epsilon(v) \operatorname{int} \mathbb{B}\} \cap \mathbb{S}^{d-1}$ form an open cover of $\mathbb{S}^{d-1}$, from which we can abstract a finite subcover, and indexing by $v_1, \ldots, v_K$ and the associated $z_1, \ldots, z_K$, we obtain

$$\inf_{v \in \mathbb{S}^{d-1}} \max_{i \leq K} \mathsf{D}\ell_{z_i}(\theta; v) = \min_{j \leq K} \inf_{\|u-v_j\|_2 < \epsilon} \max_{i \leq K} \mathsf{D}\ell_{z_i}(\theta; u) \geq \min_{j \leq K} \max_{i \leq K} \inf_{\|u-v_j\|_2 < \epsilon} \mathsf{D}\ell_{z_i}(\theta; u) > 0.$$

Written differently, we have the following strengthening of Definition 2.1.

Lemma B.1. *Let $\theta \in \operatorname{int} \Theta$ satisfy Definition 2.1. Then there exists a finite $k \in \mathbb{N}$, set $\{z_1, \ldots, z_k\} \subset \mathcal{Z}$, and $c = c(\theta) > 0$ such that for each $v \in \mathbb{S}^{d-1}$, there is z_i for which*

$$\mathsf{D}\ell_{z_i}(\theta; v) \geq c. \tag{19}$$

The next result provides a theorem of alternatives, in the vein of Farkas' lemma.

Lemma B.2. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows $a_1^T, \ldots, a_m^T$ and $u \in \mathbb{R}^n$ a vector. Then precisely one of the following two alternatives holds:*

(a) There exists $\lambda \in \mathbb{R}_+^m$, $\langle \lambda, \mathbf{1} \rangle = 1$ for which $A^T \lambda = u$

(b) There exists a vector v for which $\langle a_i - u, v\rangle > 0$ for each i.

Proof By homogeneity, part (b) is equivalent to there existing v, $\|v\|_2 \le 1$, for which $\langle a_i - u, v\rangle > 0$. Let $\Lambda = \{\lambda \in \mathbb{R}^m_+, \langle \lambda, \mathbf{1}\rangle = 1\}$. Then

$$\sup_{v\in\mathbb{B}_2^n} \inf_{\lambda\in\Lambda} v^T(A^T\lambda - u) = \inf_{\lambda\in\Lambda} \sup_{v\in\mathbb{B}_2^n} v^T(A^T\lambda - u)$$

by standard min-max duality with bilinear functions. So if condition (a) holds for some $\lambda_\star$, then obviously $\inf_{\lambda\in\Lambda}\sup_{v\in\mathbb{B}_2} v^T(A^T\lambda - u) \le \sup_{v\in\mathbb{B}_2} v^T(A^T\lambda_\star - u) = 0$, while if condition (b) held, we would have $\inf_{\lambda\in\Lambda} v^T(A^T\lambda - u) = \min_i\langle a_i - u, v\rangle > 0$, a contradiction. Conversely, if (a) fails, then $\sup_{v\in\mathbb{B}_2^n} v^T(A^T\lambda - u) = \left\|A^T\lambda - u\right\|_2 > 0$, and as Λ is compact, $\inf_{\lambda\in\Lambda}\left\|A^T\lambda - u\right\|_2 > 0$. The value of the game is thus positive, and so $\sup_{v\in\mathbb{B}_2^n}\min_i\langle a_i - u, v\rangle > 0$, i.e., (b) holds. □

Combining Lemma B.1 and Lemma B.2 will allow us to prove Lemma A.4. Take the set $\{z_i\}_{i=1}^k$ that Lemma B.1 guarantees, so that there is a $c > 0$ such that for each $v \in \mathbb{S}^{d-1}$ there exists i and $g_i \in \partial\ell_{z_i}(\theta)$ satisfying

$$\langle g_i, v\rangle \ge c.$$

We will show that for $0 \le t < c$, there is a solution to $\sum_{i=1}^m \lambda_i g_i = tu$ with $\lambda \succeq 0, \langle \mathbf{1}, \lambda\rangle = 1$, which is evidently equivalent to Lemma B.1. Take $t < c$, which in turn implies

$$\langle g_i - tu, v\rangle \ge c - t > 0.$$

In particular, for any $v \in \mathbb{S}^{d-1}$, there exist indices i and j for which

$$\langle g_i - tu, v\rangle > 0 \;\text{ and }\; \langle g_j - tu, -v\rangle > 0 \;\text{ i.e. }\; \langle g_j - tu, v\rangle < 0.$$

That is, for any $v \in \mathbb{S}^{d-1}$,

$$\max_{i\le k}\langle g_i - tu, v\rangle > 0 \;\text{ and }\; \min_{i\le k}\langle g_i - tu, v\rangle < 0.$$

Lemma B.2 thus demonstrates that $\sum_{i=1}^m \lambda_i g_i = tu$ with $\lambda \succeq 0$, $\langle \mathbf{1}, \lambda\rangle = 1$ has a solution.

B.2 Proof of Lemma 2.1

Recall the squared Hellinger distance $d^2_{\rm hel}(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$. The Hellinger and variation distances satisfy Le Cam's inequalities (Tsybakov, 2009, Lemma 2.3)

$$d^2_{\rm hel}(P,Q) \le \|P - Q\|_{\rm TV} \le d_{\rm hel}(P,Q)\sqrt{2 - d^2_{\rm hel}(P,Q)}$$

for any P, Q, and the Hellinger distance tensorizes via $d^2_{\rm hel}(P^n,Q^n) = 1-(1-d^2_{\rm hel}(P,Q))^n$. For the first statement,

$$\begin{aligned}\|P_0^n - P_1^n\|_{\rm TV} &\le d_{\rm hel}(P_0^n,P_1^n)\sqrt{2 - d^2_{\rm hel}(P_0^n,P_1^n)}\\ &= \sqrt{1-(1-d^2_{\rm hel}(P_0,P_1))^n}\sqrt{1+(1-d^2_{\rm hel}(P_0,P_1))^n}\\ &\le \sqrt{1-(1-\gamma)^n}\sqrt{1+(1-\gamma)^n} = \sqrt{1-(1-\gamma)^{2n}}\end{aligned}$$

Taking $P = P_{0,n}$ and $Q = P_{1,n}$ and setting $\gamma = \frac{a}{n}$ yields

$$\left\|P_{0,n}^n - P_{1,n}^n\right\|_{\rm TV} \le \sqrt{1 - \left(1 - \frac{a}{n}\right)^{2n}}.$$

Take $n \to \infty$.

B.3 Proof of Lemma C.3

We prove the results about M_θ^-. We know that $M_\theta^- = \inf_{z\in\mathcal{Z}} \mathsf{D}_-\ell_z(\theta)$ by Lemma C.1, part (i). We will demonstrate that for all $t > 0$, $M_\theta^- \le M_{\theta+t}^-$. If $M_{\theta+t}^- = -\infty$, then for any $M < \infty$, there exists $z \in \mathcal{Z}$ such that $\mathsf{D}_-\ell_z(\theta) \le -M$, and the monotonicity of $\theta \mapsto \mathsf{D}_-\ell_z(\theta)$ (Lemma C.1, part (iii)) implies $M_\theta^- \le \mathsf{D}_-\ell_z(\theta) \le -M$. As M was arbitrary, we have $M_\theta^- = -\infty$. If $M_{\theta+t}^- > -\infty$, then for any $\epsilon > 0$, there exists z such that $\mathsf{D}_-\ell_z(\theta + t) \le M_{\theta+t}^- + \epsilon$. For this z, $M_\theta^- \le \mathsf{D}_-\ell_z(\theta) \le \mathsf{D}_-\ell_z(\theta + t) \le M_{\theta+t}^- + \epsilon$, proving the monotonicity. The monotonicity of M_θ^+ is similar.

B.4 Proof of Lemma C.18

We first demonstrate that it is no loss of generality to assume that the losses ℓ_z are nonnegative. Because point masses $\mathbf{1}_z \in \mathcal{P}_{\rm disc}(\mathcal{Z})$ trivially, we see that Condition 2 implies that $\ell_z^\star(\Theta) > -\infty$ for all $z \in \mathcal{Z}$, as $\ell_z^\star(\Theta_0) > -\infty$ by propriety of the losses. Now we show that we may reduce to the case that the losses are non-negative. Define the rescaled losses $\overline{\ell}_z = \ell_z - \ell_z^\star(\Theta_0)$, so that $\overline{\ell}_z \ge 0$ on Θ_0, $\overline{\ell}_z$ is still z-measurable, and

$$\overline{\ell}_z(\theta) \ge -\epsilon \;\; \text{for all } \theta \in \Theta, z \in \mathcal{Z}.$$

Condition 2 includes Condition 1, so for any $\theta_0 \in \operatorname{int}\Theta$, there is a Lipschitz constant M_0 such that $M_0 \ge \sup_{g,z}\{\|g\|_2 \mid g \in \partial \ell_z(\theta_0)\}$. Thus, assuming w.l.o.g. that $\theta_0 \in \Theta_0$ by increasing the compact set Θ_0 if necessary, for any $g_0 \in \partial \ell_z(\theta_0)$ we have the lower bound

$$\ell_z(\theta) \ge \ell_z(\theta_0) + \langle g_0, \theta - \theta_0 \rangle \ge \ell_z(\theta_0) - M_0 \left\|\theta - \theta_0\right\|_2$$

so

$$\ell_z^\star(\Theta_0) \ge \ell_z(\theta_0) - M_0 \cdot \operatorname{diam}(\Theta_0).$$

Then $-\infty < \mathbb{E}_P[\ell_Z^\star(\Theta_0)] < \infty$, and so for any $\theta_0 \in \Theta_0$, $\theta \in \Theta$, and any $P \in \mathcal{P}_\ell(\mathcal{Z})$,

$$\begin{aligned} L_P(\theta_0) - L_P(\theta) &= \mathbb{E}_P\left[\ell_Z(\theta_0) - \ell_Z^\star(\Theta_0)\right] - \mathbb{E}_P\left[\ell_Z(\theta) - \ell_Z^\star(\Theta_0)\right] \\ &= \mathbb{E}_P\left[\overline{\ell}_Z(\theta_0)\right] - \mathbb{E}_P\left[\overline{\ell}_Z(\theta)\right]. \end{aligned}$$

Adding an additional $\epsilon > 0$, we see it is no loss of generality to assume that the losses are nonnegative, $\ell_z \ge 0$.

With the w.l.o.g. nonnegativity assumption, we can show the extension to Borel measures from discrete. Recall that functions f_n epi-converge to f, denoted $f_n \overset{\rm epi}{\to} f$, if whenever $\theta_n \to \theta$,

$$\liminf_n f_n(\theta_n) \ge f(\theta)$$

and for each θ, there exists a sequence $\theta_n \to \theta$ such that

$$\limsup_{n\to\infty} f_n(\theta_n) \le f(\theta).$$

Shapiro et al. (2014, Thm. 7.49) shows that $L_{P_n} \overset{\rm epi}{\to} L_P$ with probability 1 for all $P \in \mathcal{P}_\ell$, and moreover, for any compact Θ_0, they also show (Shapiro et al., 2014, Thm. 5.4) that $L_{P_n}^\star(\Theta_0) \to L_P^\star(\Theta_0)$ with probability 1. As Condition 2 holds, let $\epsilon > 0$ be arbitrary and assume Θ_0 satisfies $L_Q^\star(\Theta_0) \le L_Q^\star(\Theta) + \epsilon$ for all discrete Q. Then taking P_n to be the empirical distribution of $\{Z_i\}_{i=1}^n$ drawn i.i.d. P, which is discrete, we may assume that we have

$$L_{P_n} \overset{\rm epi}{\to} L_P \;\; \text{and} \;\; L_{P_n}^\star(\Theta_0) \to L_P^\star(\Theta_0)$$

for compact Θ_0. Because L_P is lower semicontinuous and bounded below (by the w.l.o.g. assumption that $\ell_z \ge 0$), there exists $\theta \in \Theta$ such that $L_P^\star(\Theta) \le L_P(\theta) + \epsilon$. For this θ, epi-convergence implies there exists a sequence $\theta_n \to \theta$, $\theta_n \in \Theta$, for which $\limsup_n L_{P_n}(\theta_n) \le L_P(\theta)$. Thus

$$L_P^\star(\Theta_0) = \lim_n L_{P_n}^\star(\Theta_0) \overset{(i)}{\le} \limsup_n L_{P_n}^\star(\Theta) + \epsilon \le \limsup_n L_{P_n}(\theta_n) + \epsilon \overset{(ii)}{\le} L_P(\theta) + \epsilon,$$

where inequality (i) follows by Condition 2 and inequality (ii) by $L_{P_n} \overset{\rm epi}{\to} L_P$. By assumption, $L_P(\theta) \le L_P^\star(\Theta) + \epsilon$, so

$$L_P^\star(\Theta_0) \le L_P^\star(\Theta) + 2\epsilon.$$

B.5 Proof of Lemma A.1

We recall a few results on differentiability and closedness of stochastic optimization functionals Bertsekas (1973); Shapiro et al. (2014). When $P \in \mathcal{P}_\ell$, the probability distributions making L_P well-defined over Θ, $\theta \mapsto L_P(\theta)$

is closed convex whenever $\ell_z(\cdot)$ is closed convex for each $z \in \mathcal{Z}$, and moreover, we can interchange directional differentiation and integration (see, e.g., Bertsekas (1973) or (Shapiro et al., 2014, Thm. 7.52)):

$$\mathsf{D}_+ L_P(\theta) = \int \mathsf{D}_+ \ell_z(\theta) dP(z) \quad \text{and} \quad \mathsf{D}_- L_P(\theta) = \int \mathsf{D}_- \ell_z(\theta) dP(z).$$

Recalling the definition (16) of $\theta_{\min}$ and $\theta_{\max}$ in terms of directional derivatives, we have

$$\theta_{\min} = \inf \left\{ \theta \in \Theta \mid \sup_{z \in \mathcal{Z}} \mathsf{D}_+ \ell_z(\theta) > 0 \right\} \quad \text{and} \quad \theta_{\max} = \sup \left\{ \theta \in \Theta \mid \inf_{z \in \mathcal{Z}} \mathsf{D}_- \ell_z(\theta) < 0 \right\}.$$

Thus, the swapping of integration and directional differentiation implies

$$\mathsf{D}_+ L_P(\theta_{\max}) = \int \mathsf{D}_+ \ell_z(\theta_{\max}) dP(z) \overset{(\star)}{=} \int \lim_{\theta \downarrow \theta_{\max}} \underbrace{\mathsf{D}_+ \ell_z(\theta)}_{\ge 0} dP(z) \ge 0,$$

where equality $(\star)$ follows from Lemma C.1. Similarly,

$$\mathsf{D}_- L_P(\theta_{\min}) \le 0.$$

If $\theta_{\max} = \theta_{\min}$, these equalities imply that $\theta^\star = \theta_{\max} = \theta_{\min}$ minimizes L_P, and as the choice of P was arbitrary, this implies part (i) in that case. For $\theta < \theta_{\min}$, we have $\sup_{z \in \mathcal{Z}} \mathsf{D}_+ \ell_z(\theta) \le 0$, while $\theta > \theta_{\max}$ implies $\inf_{z \in \mathcal{Z}} \mathsf{D}_- \ell_z(\theta) \ge 0$. In particular, $0 \le \mathsf{D}_- \ell_z(\theta) \le \mathsf{D}_+ \ell_z(\theta) \le 0$ for all z, that is, $\ell_z'(\theta) = 0$ for any $\theta_{\max} < \theta < \theta_{\min}$, completing the proof of part (i) once we recognize that

$$\mathsf{D}_+ \ell_z(\theta_{\max}) \le 0 = \mathsf{D}_- \ell_z(\theta) = \mathsf{D}_+ \ell_z(\theta) \le \mathsf{D}_- \ell_z(\theta_{\min})$$

for $\theta_{\max} < \theta < \theta_{\min}$.

Now we consider the cases that $\theta_{\min} \le \theta_{\max}$. Let $\Theta^\star(P) = \operatorname{argmin}_{\theta \in \Theta} L_P(\theta)$; because L_P is closed convex, its minimizers are attained (once we allow extended reals). We consider the left endpoint $\theta_{\min}$, as the arguments for the right endpoint are similar, and wish to show that $[\theta_{\min}, \infty] \cap \Theta^\star(P) \neq \emptyset$. If $\theta_{\min} = \inf\{\theta \in \Theta\}$, there is nothing to show, so assume instead that $\theta_{\min} > \inf\{\theta \in \Theta\}$, whence $-\infty < L_P(\theta_{\min}) < \infty$. Then for $\theta < \theta_{\min}$, assuming $L_P(\theta) < \infty$, we have $\sup_{z \in \mathcal{Z}} \mathsf{D}_+ \ell_z(\theta) \le 0$, and applying Lemma C.2,

$$\inf_{\theta \in \Theta} L_P(\theta) \le L_P(\theta_{\min}) = L_P(\theta) + \int_\theta^{\theta_{\min}} \underbrace{\mathsf{D}_+ L_P(t)}_{\le 0} dt \le L_P(\theta).$$

Performing the same argument for $\theta_{\max}$, we see that if $\theta \notin [\theta_{\min}, \theta_{\max}]$, then

$$L_P(\theta) \ge \inf_{\theta \in [\theta_{\min}, \theta_{\max}]} L_P(\theta),$$

yielding parts (ii) and (iii) of the lemma.

Finally, we turn to the last claim (iv). Let

$$\Theta^* := \left\{ \theta \in \Theta \mid \inf_{z \in \mathcal{Z}} \mathsf{D}_- \ell_z(\theta) < 0, \sup_{z \in \mathcal{Z}} \mathsf{D}_+ \ell_z(\theta) > 0 \right\}$$

denote the variant of the set on the right side of the claimed equality (17) whose closure we have not taken. When $\theta_{\min} < \theta_{\max}$, for any $\theta \in (\theta_{\min}, \theta_{\max})$ there exist z_0, z_1 such that $\mathsf{D}_+ \ell_{z_0}(\theta) > 0$ and $\mathsf{D}_- \ell_{z_1}(\theta) < 0$, so $\operatorname{int} \Theta_{\rm ach} \subset \Theta^*$. For the converse, observe that because $\Theta_{\rm ach}$ has interior, Θ^* is non-empty. It is also an interval (though it may be empty): if $\theta_0 < \theta_1$ and for each $i \in \{0, 1\}$, there exists $z_i^\pm$ satisfying $\mathsf{D}_- \ell_{z_i^-}(\theta_i) < 0$ and $\mathsf{D}_+ \ell_{z_i^+}(\theta_i) > 0$, then for $\theta_0 < \theta < \theta_1$, the monotonicity of subgradients (Lemma C.1, part (ii)) implies that

$$\mathsf{D}_- \ell_{z_1^-}(\theta) < 0 \le \mathsf{D}_- \ell_{z_1^-}(\theta_1) < 0 \quad \text{and} \quad \mathsf{D}_+ \ell_{z_0^+}(\theta) \ge \mathsf{D}_+ \ell_{z_0^+}(\theta_0) > 0,$$

so that $\theta \in \Theta^*$. If $\theta \in \Theta^*$, then clearly $\theta_{\min} \le \theta \le \theta_{\max}$ by construction, implying $\operatorname{cl} \Theta^* \subset \Theta_{\rm ach}$.

B.6 Proof of Lemma A.2

Let $C^\star = \cap\{C \in \mathcal{A}(\ell, \mathcal{Z}, \Theta)\}$ be the set on the right side of the equality in Lemma A.2. We first show that $\Theta_{\rm ach} \subset C^\star$. Assume w.l.o.g. that $\Theta_{\rm ach} \neq \emptyset$ and let $\theta \in \Theta_{\rm ach}$. If $\Theta_{\rm ach} = \{\theta\}$, then part (i) of Lemma A.1 gives that $\theta \in C^\star$. It is straightforward to see that if $\theta \in \operatorname{int}\Theta_{\rm ach}$, then $\theta \in C$ for each $C \in \mathcal{A}(\ell, \mathcal{Z}, \Theta)$. Indeed, by definition (8), there exist z^+ and $z^- \in \mathcal{Z}$ such that $\mathsf{D}_+\ell_{z^+}(\theta) > 0$ and $\mathsf{D}_-\ell_{z^-}(\theta) < 0$. Set p to solve $p\mathsf{D}_+\ell_{z^+}(\theta) + (1-p)\mathsf{D}_-\ell_{z^-}(\theta) = 0$, and then define the distribution $P = p\mathbf{1}_{z^+} + (1-p)\mathbf{1}_{z^-}$. Evidently, θ minimizes L_P. Because $C^\star$ is closed, it must therefore contain $\operatorname{cl}\Theta_{\rm ach}$.

Suppose $\theta \notin \Theta_{\rm ach} = [\theta_{\min}, \theta_{\max}]$, and w.l.o.g. assume $\theta < \theta_{\max}$. Then by part (ii) of Lemma A.1, $\sup\{\theta' \in \operatorname{argmin} L_P\} \geq \theta_{\min} > \theta$ for all distributions P. The closed set $C = [\theta_{\min}, \infty]$ thus belongs to $\mathcal{A}(\ell, \mathcal{Z}, \Theta)$, and $\theta \notin C$. We have shown $\Theta_{\rm ach}^c \subset (C^\star)^c$, that is, $C^\star \subset \Theta_{\rm ach}$ as desired.

B.7 Proof of Lemma A.5

If $\mathcal{P}_\ell$ contains well-defined probability measures, then standard consistency results for stochastic approximation (Shapiro et al., 2014, Thm. 5.4) show that if P_n denotes the empirical measure for $Z_i \stackrel{\rm iid}{\sim} P$, then whenever $S_0(L_P) = \operatorname{argmin}_{\theta\in\Theta} L_P(\theta)$ is non-empty and bounded, then

$$L^\star_{P_n}(\Theta) \stackrel{a.s.}{\to} L^\star_P(\Theta) \ \text{ and } \ L^\star_{P_n}(\Theta \cap H_{v,t}) \stackrel{a.s.}{\to} L^\star_{P_n}(\Theta \cap H_{v,t}).$$

In particular, if $H_{v,t}$ is unconstraining for all finitely supported measures, then for all $\epsilon > 0$, there exists a finitely supported measure Q such that

$$\left|L^\star_P(H_{v,t} \cap \Theta) - L^\star_Q(H_{v,t} \cap \Theta)\right| \leq \epsilon \ \text{ and } \ \left|L^\star_P(\Theta) - L^\star_Q(\Theta)\right| \leq \epsilon.$$

In particular,

$$|L^\star_P(H_{v,t} \cap \Theta) - L^\star_P(\Theta)| \leq 2\epsilon$$

for all $\epsilon > 0$, that is, $L^\star_P(H_{v,t} \cap \Theta) = L^\star_P(\Theta)$.

B.8 Proof of Lemma A.6

Let $E(\theta_0, \ell) \cap \Theta = \{\theta_0\}$, which is equivalent to the condition that

$$\Theta \subset (E(\theta_0, \ell) \setminus \{\theta_0\})^c = \{\theta \mid \text{there exists } z \in \mathcal{Z} \text{ s.t. } \mathsf{D}\ell_z(\theta_0; \theta - \theta_0) < 0\}.$$

For each $\theta_1 \in \Theta$, there exists $z \in \mathcal{Z}$ for which $\mathsf{D}\ell_z(\theta_0; \theta_1 - \theta_0) < 0$. As a consequence, because for any $v \in \mathbb{S}^{d-1}$ we may take $\theta_1 = \theta_0 + tv \in \Theta$ for some $t > 0$, there exists z for which $\mathsf{D}\ell_z(\theta_0; v) = \frac{1}{t}\mathsf{D}\ell_z(\theta_0; \theta_1 - \theta_0) < 0$. For this z, we obtain

$$0 > \mathsf{D}\ell_z(\theta_0; v) = \sup_{g \in \partial\ell_z(\theta_0)} \langle g, v\rangle = -\inf_{g \in \partial\ell_z(\theta_0)} \langle g, -v\rangle = -\mathsf{D}_-\ell_z(\theta_0; -v),$$

i.e., $\mathsf{D}_-\ell_z(\theta_0; -v) > 0$. So $\mathsf{D}\ell_z(\theta_0; -v) \geq \mathsf{D}_-\ell_z(\theta_0; -v) > 0$, and as $v \in \mathbb{S}^{d-1}$ was arbitrary (so we may as well have chosen $-v$), we see that for any $v \in \mathbb{S}^{d-1}$, there exists $z \in \mathcal{Z}$ for which $\mathsf{D}\ell_z(\theta_0; v) > 0$. That is, Definition 2.1 holds at θ_0.

For the converse, assume that $\theta \in \operatorname{int}\Theta$ satisfies Definition 2.1. Then for $\theta_0 \in \operatorname{int}\Theta$, $\theta_0 \neq \theta$, setting $v = (\theta_0 - \theta)/\left\|\theta - \theta_0\right\|_2$, there exists z for which $\mathsf{D}\ell_z(\theta; v) > 0$. The function $h(t) := \ell_z(t\theta_0 + (1-t)\theta)$ satisfies

$$\partial h(t) = [\mathsf{D}_-\ell_z(t\theta_0 + (1-t)\theta; \theta_0 - \theta), \mathsf{D}\ell_z(t\theta_0 + (1-t)\theta; \theta_0 - \theta)]$$

and that $\partial h(t)$ is increasing (or at least non-decreasing) in t (see (Hiriart-Urruty and Lemaréchal, 1993a, Ch. VI.2.3)). In particular, $\mathsf{D}_-\ell_z(\theta_0; \theta_0 - \theta) \geq \mathsf{D}\ell_z(\theta; \theta_0 - \theta) > 0$. Rewriting this,

$$0 < \mathsf{D}_-\ell_z(\theta_0; \theta_0 - \theta) = \inf_{g \in \partial\ell_z(\theta_0)} \langle g, \theta_0 - \theta\rangle = -\sup_{g \in \partial\ell_z(\theta_0)} \langle g, \theta - \theta_0\rangle = -\mathsf{D}\ell_z(\theta_0; \theta - \theta_0).$$

That is, $\mathsf{D}\ell_z(\theta_0; \theta - \theta_0) < 0$, and $\theta \notin E(\theta_0, \ell)$ for any $\theta_0 \neq \theta, \theta_0 \in \operatorname{int}\Theta$.

B.9 Proof of Lemma A.7

For each pair $\theta_0, \theta_1 \in \Theta$ with $\theta_1 \neq \theta_0$, there exists z such that $\ell_z(\theta_1) > \ell_z(\theta_0)$. Written differently, for each $\theta \in \operatorname{int}\Theta$ and each $v \in \mathbb{S}^{d-1}$ and $t > 0$ such that $\theta - tv \in \Theta$, there exists z such that

$$\frac{\ell_z(\theta) - \ell_z(\theta - tv)}{t} > 0.$$

Because ℓ_z is subdifferentiable at $\theta \in \operatorname{int}\Theta$, $\ell_z(\theta - tv) \geq \ell_z(\theta) - t\langle g, v\rangle$ for each $g \in \partial \ell_z(\theta)$, that is, $t\langle g, v\rangle \geq \ell_z(\theta) - \ell_z(\theta - tv)$. Rearranging the preceding display,

$$\langle g, v\rangle \geq \frac{\ell_z(\theta) - \ell_z(\theta - tv)}{t} > 0 \ \text{ for all } g \in \partial \ell_z(\theta),$$

and so $\mathsf{D}_-\ell_z(\theta; v) = \inf_{g \in \partial \ell_z(\theta)} \langle g, v\rangle > 0$.

For the converse, let θ_0 satisfy Definition 2.1 and $\theta \neq \theta_0$, $\theta \in \Theta$. Then for $v = (\theta - \theta_0)/\left\|\theta - \theta_0\right\|_2$, we take a z for which $\mathsf{D}\ell_z(\theta_0; v) > 0$, and then

$$\ell_z(\theta) = \ell_z(\theta_0) + \left\|\theta - \theta_0\right\|_2 \cdot \int_0^1 \underbrace{\mathsf{D}\ell_z(\theta_0 + t(\theta - \theta_0); v)}_{\geq \mathsf{D}\ell_z(\theta_0; v) > 0} dt \geq \ell_z(\theta_0) + \left\|\theta_0 - \theta\right\|_2 \mathsf{D}\ell_z(\theta_0; v).$$

So $\theta \notin \mathsf{domin}_\ell(\theta_0)$.

C PROOFS OF THE MAIN RESULTS

We prove our main theorems here.

C.1 Background on convex functions

We begin with a few preliminary observations on convex functions, which will be useful throughout the proofs of our results and other development; we defer proofs to appendices.

C.1.1 One-dimensional first-order behavior

In the case that $\Theta \subset \mathbb{R}$, convex functions admit substantial structure that we can leverage (e.g., Hiriart-Urruty and Lemaréchal, 1993a, Ch. 1), including extensions of derivatives and subgradients to the extended real line $[-\infty, \infty]$. Recalling the shorthand $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be a (proper) closed convex function. Then we may define the left and right derivatives (Hiriart-Urruty and Lemaréchal, 1993a, Thm. I.4.1.1)

$$\mathsf{D}_+ f(\theta) := \lim_{t \downarrow 0} \frac{f(\theta + t) - f(\theta)}{t} = \inf_{t > 0} \frac{f(\theta + t) - f(\theta)}{t}, \tag{20a}$$

$$\mathsf{D}_- f(\theta) := \lim_{t \uparrow 0} \frac{f(\theta - t) - f(\theta)}{-t} = \sup_{t > 0} \frac{f(\theta - t) - f(\theta)}{-t}. \tag{20b}$$

The following lemma (Hiriart-Urruty and Lemaréchal, 1993a, Ch. I.4) characterizes these quantities everywhere on $\operatorname{dom} f$ and relates $\mathsf{D}_\pm$ to the subdifferential of f.

Lemma C.1. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be closed convex. Then f is continuous on its domain $\operatorname{dom} f = \{\theta \in \mathbb{R} \mid f(\theta) < \infty\}$, and the following properties hold.*

(i) If $\theta \in \operatorname{int}\operatorname{dom} f$, $\partial f(\theta) = [\mathsf{D}_- f(\theta), \mathsf{D}_+ f(\theta)]$ is a non-empty compact interval.

(ii) The subdifferential ∂f and directional derivatives $\mathsf{D}_\pm$ are monotone: for $\theta_0 < \theta_1$,

$$g_0 \in \partial f(\theta_0) \text{ and } g_1 \in \partial f(\theta_1) \ \text{ implies } \ g_0 \leq g_1$$

and

$$\mathsf{D}_- f(\theta_0) \leq \mathsf{D}_+ f(\theta_0) \leq \mathsf{D}_- f(\theta_1) \leq \mathsf{D}_+ f(\theta_1).$$

(iii) The directional derivatives are continuous in that for $\theta_0 \in \operatorname{int}\operatorname{dom} f$*,*

$$\lim_{\theta \uparrow \theta_0} \mathsf{D}_- f(\theta) = \lim_{\theta \uparrow \theta_0} \mathsf{D}_+ f(\theta) = \mathsf{D}_- f(\theta_0) \quad \textit{and} \quad \lim_{\theta \downarrow \theta_0} \mathsf{D}_+ f(\theta) = \lim_{\theta \downarrow \theta_0} \mathsf{D}_- f(\theta) = \mathsf{D}_+ f(\theta_0).$$

(iv) If θ_0 *is the left endpoint of* $\operatorname{dom} f$*,* $\mathsf{D}_+ f(\theta_0) = \lim_{t \downarrow 0} \sup\{g \in \partial f(\theta_0 + t)\}$ *and the defining equality* (20a) *holds in* $\overline{\mathbb{R}}$*. (Respectively, the right endpoint and* (20b) *in* $\overline{\mathbb{R}}$*).*

(v) On each $[\theta_0, \theta_1] \subset \operatorname{dom} f$*,* f *is* $M = \max\{-\mathsf{D}_- f(\theta_0), \mathsf{D}_+ f(\theta_1)\}$*-Lipschitz on* $[\theta_0, \theta_1]$*.*

A convex f is differentiable almost everywhere on the interior of its domain, and if f is defined on an interval, it has an integral form (Hiriart-Urruty and Lemaréchal, 1993a, Remark I.4.2.5):

Lemma C.2. *If* f *is closed convex and* $[\theta_0, \theta_1] \subset \operatorname{dom} f$*, then* $f(\theta_1) - f(\theta_0) = \int_{\theta_0}^{\theta_1} \mathsf{D}_+ f(t) dt = \int_{\theta_0}^{\theta_1} \mathsf{D}_- f(t) dt$*, and for almost all* $\theta \in (\theta_0, \theta_1)$*,* $\mathsf{D}_+ f(\theta) = f'(\theta) = \mathsf{D}_- f(\theta)$*.*

We may reformulate Condition 1 more precisely in terms of the Lipschitz constants of the losses ℓ_z. Define the signed *local Lipschitz constants* at each $\theta \in \Theta$ by

$$M_\theta^+ := \sup\{g \in \partial \ell_z(\theta) \mid z \in \mathcal{Z}\} \quad \text{and} \quad M_\theta^- := \inf\{g \in \partial \ell_z(\theta) \mid z \in \mathcal{Z}\} \tag{21}$$

The monotonicity properties of subdifferentials more or less immediately implies that the Lipschitz constants $M_\theta^\pm$ are monotone (see Appendix B.3):

Lemma C.3. *The function* M_θ^- *is non-decreasing in* θ*, and* M_θ^+ *is non-decreasing in* θ*.*

In view of Lemma C.3, the signed Lipschitz constants (21) define the local Lipschitz constant

$$M_\theta = \max\{M_\theta^+, -M_\theta^-\} = \max\{|M_\theta^+|, |M_\theta^-|\},$$

and Condition 1 holds if and only if $M_\theta < \infty$ on $\operatorname{int}\Theta$.

C.1.2 Multi-dimensional first-order behavior

For $\Theta \subset \mathbb{R}^d$ with $d > 1$, convex functions admit similar monotonicity and continuity properties of the subdifferential to the one-dimensional case. Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. If $\theta \in \operatorname{int}\operatorname{dom} f$, then the subdifferential $\partial f(\theta)$ is a non-empty compact convex set (Hiriart-Urruty and Lemaréchal, 1993a, Ch. VI.1). Recalling the directional derivative $\mathsf{D} f(\theta; v) = \lim_{t \downarrow 0} \frac{f(\theta + tv) - f(\theta)}{t}$, we define

$$\mathsf{D}_- f(\theta; v) := \lim_{t \uparrow 0} \frac{f(\theta - tv) - f(\theta)}{-t} \stackrel{(\star)}{=} \inf_{g \in \partial f(\theta)} \langle g, v \rangle,$$

where equality $(\star)$ holds so long as $\partial f(\theta)$ is non-empty (Hiriart-Urruty and Lemaréchal, 1993a, Ch. VI.2.3). We can analogize Lemmas C.1 and C.2 in the multi-dimensional case as well (cf. Hiriart-Urruty and Lemaréchal, 1993a, Ch. VI):

Lemma C.4. *Let* $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ *be closed convex and proper. Then it is continuous on the interior of its domain* $\operatorname{dom} f = \{\theta \in \mathbb{R}^d \mid f(\theta) < \infty\}$*, and the following hold.*

(i) On each compact set $\Theta_0 \subset \operatorname{int}\operatorname{dom} f$*,* f *is Lipschitz with constant*

$$M(\Theta_0) := \sup_{\theta \in \Theta_0} \sup_{g \in \partial f(\theta)} \|g\|_2 .$$

(ii) The subdifferentials are monotone in that whenever $\partial f(\theta_0)$ *and* $\partial f(\theta_1)$ *are non-empty,*

$$g_i \in \partial f(\theta_i) \quad \textit{implies} \quad \langle g_1 - g_0, \theta_1 - \theta_0 \rangle \geq 0.$$

(iii) The subdifferential mapping is outer semicontinuous: if $\theta_0 \in \operatorname{int}\operatorname{dom} f$*, then for each* $\epsilon > 0$ *there exists* $\delta > 0$ *such that*

$$\|\theta - \theta_0\|_2 < \delta \quad \textit{implies} \quad \partial f(\theta) \subset \partial f(\theta_0) + \epsilon \mathbb{B}_2.$$

In view of Lemma C.4, we can thus generally define the local Lipschitz constants

$$M_\theta := \sup\{\|g\|_2 \mid z \in \mathcal{Z}, g \in \partial \ell_z(\theta)\}, \tag{22}$$

which coincide with the one-dimensional case (21).

C.1.3 Minimizers, conjugacy, and coercivity

Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Recall that f is coercive if $\lim_{\|\theta\| \to \infty} f(\theta) = +\infty$, which implies that all sublevel sets $\{\theta \mid f(\theta) \le c\}$ are convex and compact. For closed convex f, coercivity relates to "derivatives of f at infinity," which in turn connect to the existence of minimizing sets. To develop this idea, fix $\theta_0 \in \operatorname{dom} f$, and recall (Hiriart-Urruty and Lemaréchal, 1993a, Def. IV.3.2.3) the recession function

$$f'_\infty(v) = \lim_{t \to \infty} \frac{1}{t}(f(\theta_0 + tv) - f(\theta_0)) = \sup_{t>0} \frac{f(\theta_0 + tv) - f(\theta_0)}{t},$$

which is positively homogeneous, closed convex in v, and independent of the choice θ_0. Similarly, for a convex set C, any $\theta_0 \in C$ defines the *recession cone* (Hiriart-Urruty and Lemaréchal, 1993a, Def. III.2.2.2)

$$C_\infty \coloneqq \{v \mid \theta_0 + tv \in C \text{ for } t \ge 0\},$$

which again is independent of the choice θ_0.

Coercivity and the positivity of recession functions coincide for closed convex f:

Lemma C.5. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then f is coercive if and only if $f'_\infty(v) > 0$ for all $v \neq 0$, that is, it is coercive on each line.*

Proof If f is coercive, it is clear that $f'_\infty(v) > 0$ for all $v \in \mathbb{S}^{d-1}$. For the converse, assume that $f'_\infty(v) > 0$ for all $v \in \mathbb{S}^{d-1}$. If f were not coercive, then for some $c < \infty$ the sublevel set $S \coloneqq \{\theta \mid f(\theta) \le c\}$ must be unbounded, and as this set is closed convex, it must thus have a recession direction $v \in S_\infty$ with $v \neq 0$ (cf. (Hiriart-Urruty and Lemaréchal, 1993a, Proposition III.2.2.3)). But then $f(\theta_0 + tv) \le c$ for all t implies $\sup_{t>0} f(\theta_0 + tv) \le c$ and so $f'_\infty(v) = 0$, a contradiction. □

Whenever $f^\star = \inf_\theta f(\theta) > -\infty$, for $\epsilon \ge 0$ let

$$S_\epsilon(f) \coloneqq \{\theta \mid f(\theta) \le f^\star + \epsilon\}$$

denote its ϵ-sub-optimality set, which is always closed convex, and it is non-empty for $\epsilon > 0$. Then Lemma C.5 immediately implies the following equivalence result.

Lemma C.6. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be proper closed convex. Then $S_0(f) = \operatorname{argmin} f$ is compact and non-empty if and only if f is coercive if and only if $f'_\infty(v) > 0$ for each $v \neq 0$.*

Optimality and subdifferentials connect via convex conjugates, including in cases where f may take on infinite values (Hiriart-Urruty and Lemaréchal, 1993b, Ch. X). Recall the conjugate $f^*(v) \coloneqq \sup_\theta\{\langle v, \theta\rangle - f(\theta)\}$. If $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is closed convex, then

$$v \in \partial f(\theta) \text{ if and only if } \theta \in \partial f^*(v) \text{ if and only if } f(\theta) + f^*(v) = \langle \theta, v\rangle, \tag{23}$$

and so

$$\operatorname*{argmin}_\theta \{f(\theta) - \langle v, \theta\rangle\} = \partial f^*(v).$$

The equality (23) extends the subdifferential to arbitrary closed convex f, so we take it as the definition of ∂f. Recalling the support function $\sigma_C(v) \coloneqq \sup_{\theta \in C}\langle v, \theta\rangle$ of a set C, we also have (Rockafellar, 1970, Thm. 13.3) that

$$\sigma_{\operatorname{dom} f^*}(v) = f'_\infty(v),$$

and therefore $0 \in \operatorname{int} \operatorname{dom} f^*$ if and only if $f'_\infty(v) > 0$ for each $v \neq 0$ (cf. (Rockafellar, 1970, Thm. 13.1)). Thus, when f is coercive, or, equivalently, the set of minimizers $S_0(f) \coloneqq \{\theta \mid f(\theta) = f^\star\}$ is compact (Lemma C.6), f^* is finite in a neighborhood of 0, and then $\partial f^*(v)$ is the usual subdifferential for v near 0. The outer semi-continuity of the sub-differential of finite convex functions (Lemma C.4, part (iii)) implies that for any $\epsilon > 0$ there therefore exists $\delta > 0$ such that $\partial f^*(v) \subset \partial f^*(0) + \epsilon\mathbb{B}$ if $\|v\|_2 \le \delta$.

As a consequence of these calculations, whenever f is coercive, for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$\bigcup_{\|v\|_2 \le \delta} \partial f^*(v) \subset S_0(f) + \epsilon\mathbb{B}.$$

Inverting this inclusion, we obtain the following proposition.

Proposition 4. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a closed convex function and assume that $S_0(f) = \operatorname{argmin} f$ is compact. Then for all $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\bigcup_{\theta \in S_0(f)+\epsilon\mathbb{B}} \partial f(\theta) \supset \delta\mathbb{B}.$$

Nothing in Proposition 4 assumes that $\operatorname{dom} f = \mathbb{R}^d$, so that it allows for constraints. Most saliently for us, let $\mathbb{I}_\Theta(\theta) = 0$ for $\theta \in \Theta$ and $+\infty$ denote the convex indicator of the set Θ. Then $L_Q + \mathbb{I}_\Theta$ is closed convex for any finitely supported Q. In particular, Lemma A.3 follows as an immediate consequence of Proposition 4, where the conclusion that $\partial(L_Q + \mathbb{I}_\Theta)(\theta)$ is non-empty is one of the results.

C.2 Proof of Theorem 1

We follow the outline in Section 2.4 to prove Theorem 1 here.

C.2.1 Proof of Theorem 1: the one-dimensional case

Constructing well-separated functions. Without loss of generality (by symmetry), consider the case that the lower Lipschitz constant (21) satisfies $M_\theta^- = -\infty$ for some $\theta \in \operatorname{int}\Theta$. Choose $\theta_0 \in \operatorname{int}\Theta$ and $\delta > 0$ to satisfy $\theta_0 < \theta_0 + \delta < \theta$, so that $M_{\theta_0}^- = M_{\theta_0+\delta}^- = -\infty$ by Lemma C.3. We first "place" the gradients and minimizers (recall Fig. 1). By Assumption 1 that minimizers are achievable, there exists z^+ such that $\mathsf{D}_+\ell_{z^+}(\theta_0) = \sup\{\partial\ell_{z^+}(\theta_0)\} > 0$ and $\mathsf{D}_+\ell_{z^+}(\theta_0) \le \mathsf{D}_+\ell_{z^+}(\theta_0+\delta) < \infty$. We use the shorthands $a = \ell'_{z^+}(\theta_0) > 0$ and $b = \ell'_{z^+}(\theta_0+\delta) > 0$ for some strictly positive elements of the subdifferentials $\partial\ell_{z^+}(\theta_0)$ and $\partial\ell_{z^+}(\theta_0+\delta)$, noting that we may treat them as constants—they do not depend on the sample size n or any other distribution or sample-dependent quantities.

For any $M < \infty$, there exists z_M such that

$$\mathsf{D}_+\ell_{z_M}(\theta_0+\delta) \le -M.$$

Now, define the point distributions

$$P_0 \coloneqq \mathbf{1}_{z^+} \quad \text{and} \quad P_1 \coloneqq \left(1 - \frac{1}{n^2}\right)\mathbf{1}_{z^+} + \frac{1}{n^2}\mathbf{1}_{z_M}.$$

With these choices, observe that there is a subderivative $L'_{P_0}(\theta_0) \in \partial L_{P_0}(\theta_0)$ satisfying

$$L'_{P_0}(\theta_0) = \ell'_{z^+}(\theta_0) = a > 0.$$

Similarly, for $\epsilon = \frac{1}{n^2}$,

$$\begin{aligned} L'_{P_1}(\theta_0+\delta) &= \ell'_{z^+}(\theta_0+\delta) + \epsilon\left[\ell'_{z_M}(\theta_0+\delta) - \ell'_{z^+}(\theta_0+\delta)\right] \\ &\le (1-\epsilon)b - \epsilon M. \end{aligned}$$

As M was arbitrary, we may choose $M = n^3$, so that $L'_{P_1}(\theta_0+\delta) \le -n/2$ for large n.

From these calculations (and recall Fig. 1), we see that for all $t \ge 0$,

$$\begin{aligned} L_{P_0}(\theta_0+t) - L_{P_0}^\star \ge L_{P_0}(\theta_0+t) - L_{P_0}(\theta_0) \ge at \quad \text{while} \\ L_{P_1}(\theta_0+\delta-t) - L_{P_1}^\star \ge L_{P_1}(\theta_0+\delta-t) - L_{P_1}(\theta_0+\delta) \ge \frac{nt}{2}. \end{aligned}$$

So at least one of the population losses L_{P_0} or L_{P_1} must be nontrivially larger than its infimum, and in particular,

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) \ge \inf_t \max\left\{at, \frac{n}{2}(\delta - t)\right\} = \frac{an\delta/2}{a+n/2} = \frac{an\delta}{2a+n} \ge a\delta. \tag{24}$$

The testing guarantee. With the optimization separation (24), we can now get the desired lower bound. By Lemma 2.1, for $\epsilon = \frac{1}{n^2}$, $\|P_0 - P_1\|_{\rm TV} \le \epsilon$ and so

$$\|P_0^n - P_1^n\|_{\rm TV} \le \sqrt{1 - (1 - n^{-2})^n} = \frac{(1 + o(1))}{\sqrt{n}}.$$

That $\mathfrak{M}_n^{\rm low}$ is non-increasing in n thus gives minimax lower bound

$$\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) \ge \frac{d_{\sf opt}(L_{P_0}, L_{P_1})}{2} (1 - \|P_0^n - P_1^n\|_{\rm TV}) \ge \frac{a\delta}{2},$$

valid for all n. Of course, as we note in the beginning of the proof, $\delta > 0$ is fixed and independent of n, as is $a = \ell'_{z^+}(\theta_0) > 0$.

C.2.2 Proof of Theorem 1: the general case

We follow the outline of the proof of the one-dimensional case and that present in Section 2.4. We find two well-separated finitely supported losses by "placing" gradients in directions that most separate the functions according to the optimization distance (12). When Condition 1 fails, the definition (22) of the Lipschitz constant M_θ shows that there is some compact set $\Theta_0 \subset \operatorname{int} \Theta$ for which $\sup_{\theta \in \Theta_0} M_\theta = +\infty$. Taking a subsequence if necessary, we may find a convergent sequence $\theta_n \to \theta_0 \in \Theta_0$ satisfying $\lim_n M_{\theta_n} = +\infty$. In particular, we have a point θ_0 where for each $\epsilon > 0$,

$$\sup_{z \in \mathcal{Z}} \sup_{\|\theta - \theta_0\|_2 < \epsilon} \sup_{g \in \partial \ell_z(\theta)} \|g\|_2 = +\infty.$$

Then for any $\epsilon > 0$ and $m < \infty$, we may find z_m^* and θ_m with $\|\theta_m - \theta_0\|_2 < \epsilon$ satisfying

$$\sup_{g \in \partial \ell_{z_m^*}(\theta_m)} \|g\|_2 = m$$

(if $\|g\|_2 > m$, we may simply increase $m \uparrow \|g\|_2$, so assuming equality is no loss of generality). Because $\partial \ell_{z_m^*}(\theta_m)$ is compact, the supremum is attained at some vector $g_m \in \partial \ell_{z_m^*}(\theta_m)$, and without loss of generality, again moving to a subsequence if necessary, we assume that $v_m := g_m / \|g_m\|_2 \to v \in \mathbb{S}^{d-1}$.

We use this direction and the directability assumptions in the definition (8) of the achievable set to construct losses whose gradients point in opposite directions, exactly as in the one-dimensional case and in the "gradient placing" construction in Figure 1 and Sec. 2.4. The following technical lemma, whose proof we defer to Section C.2.3, provides the key.

Lemma C.7. *Let Assumption 1 hold and $\theta_0 \in \operatorname{int}\Theta$. Then for all $v \in \mathbb{S}^{d-1}$ there exist $\alpha_0 > 0$ and $\beta_0 > 0$ such that, there exists a convex compact set $K \subset \Theta$ such that $\langle v, \theta_0 \rangle + \alpha_0 \le \inf_{\theta \in K} \langle v, \theta \rangle$, and for each $u \in \mathbb{S}^{d-1}$, there exists a finitely supported distribution Q such that*

$$\beta u \in \partial L_Q(\theta) \quad \textit{for some } \theta \in K \textit{ and } \beta \ge \beta_0.$$

In particular, the point θ may be taken of the form $\theta = \theta_0 + \alpha v + v_\perp$, where $\langle v_\perp, v \rangle = 0$, $\|v_\perp\|_2 \le \frac{1}{\alpha_0}$, $\alpha_0 \le \alpha \le \frac{1}{\alpha_0}$, and $\beta \ge \beta_0$.

Lemma C.7 guarantees that we can find a finitely supported distribution Q with

$$\theta_1 = \theta_0 + \alpha v + v_\perp, \quad \text{and} \quad -\beta v_m \in \partial L_Q(\theta_1),$$

where $v_\perp$ is finite, $\langle v_\perp, v \rangle = 0$, and $\alpha \ge \alpha_0 > 0$, $\beta \ge \beta_0 > 0$. Therefore

$$L_Q(\theta) \ge L_Q(\theta_1) - \beta \langle v_m, \theta - \theta_1 \rangle.$$

Assume that m is large enough that $\|v_m - v\|_2 \le \epsilon$; consider the point $\overline{\theta} = \frac{1}{2}(\theta_m + \theta_1)$ halfway between θ_1 and

θ_m. If $\langle \theta, v_m \rangle \le \langle \overline{\theta}, v_m \rangle$, we have

$$\begin{aligned}
L_Q(\theta) \ge L_Q(\theta_1) - \beta\langle v_m, \theta - \theta_1\rangle &\ge L_Q(\theta_1) - \beta\langle v_m, \overline{\theta} - \theta_1\rangle \\
&= L_Q(\theta_1) - \frac{\beta}{2}\langle v_m, \theta_m - \theta_1\rangle \\
&= L_Q(\theta_1) - \frac{\beta}{2}\langle v, \theta_0 - \theta_1\rangle - \frac{\beta}{2}\langle v_m - v, \theta_0 - \theta_1\rangle - \frac{\beta}{2}\langle v_m, \theta_m - \theta_0\rangle \\
&= L_Q(\theta_1) + \frac{\beta\alpha}{2} + \frac{\beta}{2}\langle v_m - v, \alpha v + v_\perp\rangle - \frac{\beta}{2}\langle v_m, \theta_m - \theta_0\rangle \\
&> L_Q(\theta_1) + \frac{\beta\alpha}{2} - \frac{\beta(\alpha\epsilon + \epsilon\left\|v_\perp\right\|_2)}{2} - \frac{\beta\epsilon}{2} \\
&= L_Q(\theta_1) + \frac{\beta(\alpha - \alpha\epsilon - \epsilon(1 + \left\|v_\perp\right\|_2))}{2}
\end{aligned}$$

by our choice $\theta_1 = \theta_0 + \alpha v + v_\perp$ and because $\left\|v\right\|_2 = 1$ and $\left\|\theta_m - \theta_0\right\|_2 < \epsilon$. On the other hand, if $\langle \theta, v_m \rangle \ge \langle \overline{\theta}, v_m \rangle$, then we use $\left\|g_m\right\|_2 = m$ to obtain

$$\begin{aligned}
\ell_{z_m^*}(\theta) \ge \ell_{z_m^*}(\theta_m) + m\langle g_m/m, \theta - \theta_m\rangle &= \ell_{z_m^*}(\theta_m) + m\langle v_m, \theta - \theta_m\rangle \\
&\ge \ell_{z_m^*}(\theta_m) + m\langle v_m, \overline{\theta} - \theta_m\rangle \\
&= \ell_{z_m^*}(\theta_m) + \frac{m}{2}\langle v_m, \theta_1 - \theta_m\rangle \\
&= \ell_{z_m^*}(\theta_m) + \frac{m}{2}\left(\langle v_m, \alpha v + v_\perp\rangle + \langle v_m, \theta_0 - \theta_m\rangle\right) \\
&\ge \ell_{z_m^*}(\theta_m) + \frac{m}{2}\left(\alpha(1-\epsilon) - \epsilon(1 + \left\|v_\perp\right\|_2)\right).
\end{aligned}$$

With these choices, assume $\epsilon > 0$ is small enough that $\alpha(1-\epsilon) - \epsilon(1 + \left\|v_\perp\right\|_2) > \frac{\alpha}{2}$ and define the two probability distributions

$$P_0 := \left(1 - \frac{1}{n}\right) Q + \frac{1}{n}\mathbf{1}_{z_m^*} \quad \text{and} \quad P_1 := Q.$$

Then

$$L_{P_1}(\theta) \ge L_{P_1}^\star + \frac{\beta\alpha}{4} \quad \text{if} \quad \langle v_m, \theta\rangle \le \langle v_m, \overline{\theta}\rangle.$$

Consider the converse case that $\langle v_m, \theta\rangle \ge \langle v_m, \overline{\theta}\rangle$. Because Q is finitely supported and the losses are proper, $L_Q^\star > -\infty$. So if $\langle v_m, \theta\rangle \ge \langle v_m, \overline{\theta}\rangle$, then

$$\begin{aligned}
L_{P_0}(\theta) = \left(1 - \frac{1}{n}\right) L_Q(\theta) + \frac{1}{n}\ell_{z_m^*}(\theta) &\ge \left(1 - \frac{1}{n}\right) L_Q^\star + \frac{1}{n}\ell_{z_m^*}(\theta_m) + \frac{m\alpha}{4n} \\
&\ge L_{P_0}(\theta_m) + \frac{m\alpha}{4n} + \left(1 - \frac{1}{n}\right)\left(L_Q^\star - L_Q(\theta_m)\right).
\end{aligned}$$

Because Q has finite support, it is Lipschitz in any compact neighborhood of θ_0, and so $L_Q^\star - L_Q(\theta_m)$ is uniformly bounded for all m. The choice of $m < \infty$ was otherwise arbitrary, so we may take it large enough that

$$L_{P_0}(\theta) \ge L_{P_0}^\star + 1 \quad \text{if} \quad \langle v_m, \theta\rangle \ge \langle v_m, \overline{\theta}\rangle.$$

Combining the two calculations, we see that for any θ,

$$L_{P_0}(\theta) + L_{P_1}(\theta) - L_{P_0}^\star - L_{P_1}^\star \ge \min\left\{\frac{\beta\alpha}{2}, 1\right\}.$$

That is, for some (problem-dependent) constant $c > 0$, independent of the sample size n, we have constructed P_0 and P_1 so that

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) \ge c,$$

while Lemma 2.1 yields

$$\left\|P_0^n - P_1^n\right\|_{\mathrm{TV}} \le (1 + o_n(1))\sqrt{1 - e^{-1}}\sqrt{1 + e^{-1}} \to \sqrt{1 - e^{-2}}.$$

The reduction from optimization to testing (13) then gives the theorem.

C.2.3 Proof of Lemma C.7

Let $\theta_0 \in \operatorname{int}\Theta$, $v \in \mathbb{S}^{d-1}$, and $t = \langle v, \theta_0 \rangle$. Then $H_{v,t} \cap \Theta$ and $H_{v,t}^c \cap \Theta$ are each convex sets with interiors.

Assume for the sake of contradiction that for all $\alpha_0 > 0$ with $H_{v,t+\alpha_0}^c \cap \operatorname{int}\Theta \neq \emptyset$ and all directable sets C, there is $\theta \in C$ with $\langle \theta, v \rangle \le t + \alpha_0$. Then

$$H_{v,t+\alpha_0} \cap C \neq \emptyset$$

for all directable sets C, and so $(v, t+\alpha_0) \in \mathcal{U}$, that is, the pair $(v, t+\alpha_0)$ is unconstraining. Because $H_{v,t+\alpha_0}^c \cap \operatorname{int}\Theta \neq \emptyset$, we have $H_{v,t+\alpha_0} \cap \Theta \subsetneq \Theta$, contradicting Assumption 1. Thus there exists some $\alpha_0 > 0$ and directable set C such that $C \subset H_{v,t+\alpha_0}^c$.

By the Definition 2.2 of directability, for all $\epsilon > 0$ there exists a compact convex C_ϵ between C and Θ, i.e., $C \subset C_\epsilon \subset \Theta$, with $\operatorname{dist}(\theta, C) \le \epsilon$ for all $\theta \in C_\epsilon$, and a finite collection $\{z_i\}_{i=1}^k$ and $\beta_0 > 0$ such that

$$\bigcup_{Q \in \mathcal{P}(\{z_i\}_{i=1}^k)} \bigcup_{\theta \in C_\epsilon} \partial \left(L_Q + \mathbb{I}_\Theta \right)(\theta) \supset \beta_0 \mathbb{B}.$$

Because $\alpha_0 > 0$, we can certainly take ϵ small enough that $C + \epsilon\mathbb{B} \subset H_{v,t+\alpha_0/2}^c$, and taking $K = C_\epsilon$ proves the first statement of the lemma. Every element of $\Theta \cap H_{v,t+\alpha_0/2}^c$ has the form $\theta_0 + \alpha v + v_\perp$ for some $\alpha \ge \alpha_0/2$ and $\langle v_\perp, v \rangle = 0$. That we may take $\|v_\perp\|_2$ bounded and assume $\alpha \le \frac{1}{\alpha_0}$ follows because C_ϵ is compact, so that decreasing α_0 does not change the conclusions of the lemma.

C.3 Proof of Theorem 2

If $M \coloneqq \sup_{\theta \in \Theta} M_\theta < \infty$, the result is trivial: the output $\widehat{\theta}_n$ the average of n steps of the stochastic subgradient method achieves risk

$$\mathbb{E}[L_P(\widehat{\theta}_n) - L_P^\star] \le O(1) \frac{M \operatorname{diam}(\Theta)}{\sqrt{n}} \tag{25}$$

with an appropriate stepsize Nemirovski et al. (2009). We now handle the case that the Lipschitz constants (22) can explode on the boundaries, so that $\sup_{\theta \in \Theta} M_\theta = \infty$, while Condition 1 holds.

We first argue that no loss can decrease too much on a compact set.

Lemma C.8. *Let $\ell : \operatorname{int}\Theta \to \mathbb{R}$ be convex, $\theta_0 \in \operatorname{int}\Theta$, and $M_0 \coloneqq \sup_{g \in \partial\ell(\theta_0)} \|g\|_2 < \infty$. Then*

$$\ell(\theta) \ge \ell(\theta_0) - M_0 \|\theta - \theta_0\|_2 .$$

Proof The one-dimensional convex function $h(t) \coloneqq \ell(t\theta + (1-t)\theta_0)$ satisfies $h(1) = h(0) + \int_0^1 h'(t)dt$, because it is a.e. differentiable, and $h'(t) = \mathsf{D}\ell(\theta_0 + t(\theta - \theta_0); \theta - \theta_0)$ at all points where the derivative exists. Because the directional derivative is monotone, so that $\mathsf{D}\ell(\theta + tv; v) \ge \mathsf{D}\ell(\theta; v)$ for all $t \ge 0$, and $|\mathsf{D}\ell(\theta_0; v)| \le M_0 \|v\|_2$, we therefore have

$$\begin{aligned}
\ell(\theta) &= \ell(\theta_0) + \int_0^1 \mathsf{D}\ell(\theta_0 + t(\theta - \theta_0); \theta - \theta_0)dt \\
&\ge \ell(\theta_0) + \int_0^1 \mathsf{D}\ell(\theta_0; \theta - \theta_0)dt \ge \ell(\theta_0) - \int_0^1 M_0 \|\theta - \theta_0\|_2 \, dt = \ell(\theta_0) - M_0 \|\theta - \theta_0\|_2
\end{aligned}$$

as desired. ∎

Let $\theta_0 \in \operatorname{int}\Theta$, and let $M_{\theta_0} = \sup_{z \in \mathcal{Z}} \|\partial\ell_z(\theta_0)\|_2$ be the Lipschitz constant (22). Fix $\theta \in \Theta$, let $v = \theta - \theta_0$, and define $\theta_t = \theta_0 + tv$ for $t \in [0, 1)$. Then Lemma C.8 shows that

$$\ell_z(\theta) \ge \ell_z(\theta_t) - M_{\theta_0} \cdot (1-t) \|\theta - \theta_0\|_2 .$$

Defining the "star" interior of Θ for $\delta \in [0, 1]$ by

$$\Theta_\delta \coloneqq \{\theta_0 + t(\theta - \theta_0) \mid 0 \le t \le 1 - \delta, \theta \in \Theta\} = \operatorname{Conv}\{\theta_0, (1-\delta)\Theta\}$$

we see immediately that $\operatorname{int} \Theta \supset \Theta_\delta$ and Θ_δ is compact convex. Then for any $\delta \in [0,1]$ and $\theta \in \Theta$, we obtain

$$\ell_z(\theta) \ge \inf_{\theta' \in \Theta_\delta} \ell_z(\theta') - \delta M_{\theta_0} \operatorname{diam}(\Theta).$$

Without providing any concrete analytical bounds (which depend on the structure of Θ), it is apparent that for any distribution P on $\mathcal{Z}$,

$$\inf_{\theta \in \Theta_\delta} L_P(\theta) \le \inf_{\theta \in \Theta} L_P(\theta) + \delta M_{\theta_0} \operatorname{diam}(\Theta).$$

In particular, letting $M(\Theta_\delta) = \sup_{\theta \in \Theta_\delta} M_\theta$ and performing the stochastic gradient method for n steps on the set Θ_δ gives

$$\mathbb{E}\left[L_P(\widehat{\theta}_n) - L_P^\star\right] \le \delta M_{\theta_0} \operatorname{diam}(\Theta) + \frac{M(\Theta_\delta) \operatorname{diam}(\Theta)}{\sqrt{n}},$$

via the bound (25). Take $\delta = \delta_n \to 0$ slowly enough that the right side tends to zero.

C.4 Proof of Proposition 1

Without loss of generality we assume $r(1) = 1$, as otherwise we apply the same argument to the normalized rate function $\bar{r}(x) = r(x)/r(1)$. The rate function has a continuous increasing inverse $r^{-1} : [1,\infty) \to [1,\infty)$, and we define the losses $\ell_z : [0,1] \to \overline{\mathbb{R}}$ for $z \in \{0,1\}$ by

$$\ell_z(\theta) \coloneqq \theta + z \int_\theta^1 r^{-1}(1/t) dt,$$

where we define $\ell_1(0) = +\infty$. The convexity of ℓ_z follows immediately, as

$$\ell_z'(\theta) = 1 - z r^{-1}(1/\theta) \quad \text{and} \quad \ell_z''(\theta) = \frac{z}{\theta^2} \left(r^{-1}\right)'(1/\theta) \ge 0$$

(where we tacitly use that $(r^{-1})'$ exists almost everywhere). For $\delta \in (0, 1/2)$ let us now define distributions

$$P_0 \coloneqq \mathbf{1}_0 \quad \text{and} \quad P_\delta \coloneqq (1-\delta)\,\mathbf{1}_0 + \delta \mathbf{1}_1.$$

Under P_0 we have $L_{P_0}(\theta) = \theta$, which satisfies $L_{P_0}(0) = \inf_{\theta \in [0,1]} L_{P_0}(\theta) = 0$, while under P_δ

$$L_{P_\delta}(\theta) = \theta + \delta \int_\theta^1 r^{-1}(1/x) dx,$$

has minimizer $\theta_\delta = \frac{1}{r(1/\delta)}$.

We lower bound the optimization distance via

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_\delta}) \ge \inf_{\theta \in [0,1]} \frac{1}{2} \left(L_{P_0}(\theta) + L_{P_\delta}(\theta) - 0 - L_{P_\delta}(\theta_\delta)\right).$$

Evidently $\theta_{\delta/2}$ minimizes $L_{P_0}(\theta) + L_{P_\delta}(\theta)$, so that

$$\begin{aligned} d_{\mathsf{opt}}(L_{P_0}, L_{P_\delta}) &\ge \frac{1}{2}\left(\frac{2}{r(2/\delta)} - \frac{1}{r(1/\delta)} + \delta \int_{1/r(2/\delta)}^{1/r(1/\delta)} r^{-1}(1/x) dx\right) \\ &\ge \frac{1}{2}\left(\frac{2}{r(2/\delta)} - \frac{1}{r(1/\delta)} + \delta\left(\frac{1}{r(1/\delta)} - \frac{1}{r(2/\delta)}\right) r^{-1}(r(1/\delta))\right) \\ &= \frac{1}{2r(2/\delta)}, \end{aligned}$$

where we have again used that r^{-1} is increasing. Choosing $\delta = \frac{1}{n}$ and applying Lemma 2.1, we obtain

$$\mathfrak{M}_n^{\mathrm{low}}(\ell, \mathcal{Z}, \Theta) \ge \frac{1}{4r(2n)}\left(1 - \sqrt{1 - e^{-2}}\right).$$

Set the numerical constant $c = \frac{1 - \sqrt{1 - e^{-2}}}{4}$.

C.5 Proof of Theorem 3: the one-dimensional case

We begin with a few preliminary results before moving to the proof proper of Theorem 3 when $\Theta \subset \mathbb{R}$. Throughout, we assume Condition 1 holds, because otherwise, Theorem 1 gives the result. First, we show the equivalence between the conditions (10) and (11).

Lemma C.9. *Assume that* $\inf \Theta = -\infty$. *Then the following two statements are equivalent:*

(i) For all $\epsilon > 0$, *there exists* $t > -\infty$ *such that* $\inf_{\theta \ge t} [L_Q(\theta) - L_Q^\star] \le \epsilon$ *for all* $Q \in \mathcal{P}_{\rm disc}(\mathcal{Z})$.

(ii) For all $\epsilon > 0$, *there exists* $t > -\infty$ *such that* $\inf_{\theta \ge t} [\ell_z(\theta) - \ell_z^\star] \le \epsilon$ *for all* $z \in \mathcal{Z}$.

Proof Clearly (i) implies (ii). For the converse, let $t_0 > -\infty$ satisfy $\inf_{\theta \ge t_0} \ell_z(\theta) \le \ell_z^\star + \epsilon$, and consider any distribution P for which L_P is well-defined (9). Let $\theta^\star(P) \in \operatorname{argmin}_\theta L_P(\theta)$, where we tacitly let $\theta^\star(P)$ be in the extended reals $[-\infty, \infty]$. If $\theta^\star = \theta^\star(P) < t_0$, then

$$\mathbb{E}_P[\ell_Z(t_0)] = L_P(\theta^\star) + \mathbb{E}_P[\ell_Z(t_0) - \ell_Z(\theta^\star)]$$

so by assumption (11), the claim (i) holds:

$$\ell_z(t_0) - \ell_z(\theta^\star) \le \begin{cases} \ell_z(t_0) - \ell_z^\star \le \epsilon & \text{if } \mathsf{D}_+\ell_z(t_0) \ge 0 \\ 0 & \text{if } \mathsf{D}_+\ell_z(t_0) < 0. \end{cases}$$

This gives $\mathbb{E}_P[\ell_Z(t_0)] \le L_P(\theta^\star) + \epsilon$ as desired. □

C.5.1 Eliminating infinite losses

We begin the proof of Theorem 3 (when $d = 1$) by eliminating issues that arise if losses can take on infinite (or asymptotically infinite) values, showing that $\mathfrak{M}_n^{\rm low} = +\infty$ in these cases.

Lemma C.10. *If there exists* $z \in \mathcal{Z}$ *such that* $\ell_z^\star(\Theta) = -\infty$, *then* $\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) = +\infty$.

Proof Let z be such that $\ell_z^\star(\Theta) = -\infty$, and take any $\theta_0 \in \operatorname{int}\Theta$ at which ℓ_z is differentiable (this occurs for a.e. θ_0). Because ℓ_z is proper, w.l.o.g., we may assume $\lim_{t \downarrow -\infty} \ell_z(t) = -\infty$, and so $\ell_z'(\theta_0) > 0$ for all θ_0. Now, let $\theta_1 > \theta_0$, and find z_1 such that $\mathsf{D}_+\ell_{z_1}(\theta_1) < 0$, which exists by Assumption 1; let β solve $-\beta\ell_z'(\theta_0) \in \partial\ell_{z_1}(\theta_1)$, so that $\beta > 0$. Then

$$\begin{aligned} \ell_{z_1}(\theta) + \beta\ell_z(\theta) &\ge \ell_{z_1}(\theta_1) - \beta\ell_z'(\theta_0)(\theta - \theta_1) + \beta\ell_z(\theta_0) + \beta\ell_z'(\theta_0)(\theta - \theta_0) \\ &= \ell_{z_1}(\theta_1) + \beta\ell_z(\theta_0) + \beta\ell_z'(\theta_0)(\theta_1 - \theta_0) > \ell_{z_1}(\theta_1) + \beta\ell_z(\theta_0). \end{aligned}$$

Thus

$$\inf_\theta [\ell_{z_1}(\theta) + \beta\ell_z(\theta)] > \ell_{z_1}(\theta_1) + \beta\ell_z(\theta_0) > -\infty,$$

and so defining the probability distributions

$$P_0 := \mathbf{1}_z \quad \text{and} \quad P_1 := \frac{1}{1+\beta}\mathbf{1}_{z_1} + \frac{\beta}{1+\beta}\mathbf{1}_z$$

we obtain

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) = +\infty.$$

Using the shorthand $\epsilon = 1 - \frac{1}{1+\beta} = \frac{\beta}{1+\beta}$, we have $\|P_0 - P_1\|_{\rm TV} \le 1 - \epsilon$, and Lemma 2.1 implies $\|P_0^n - P_1^n\|_{\rm TV} \le \sqrt{1 - \epsilon^{2n}} < 1$. Inequality (13) shows that $\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) = +\infty$. □

Lemma C.11. *Let Condition 1 hold. If for some* $\theta_0 \in \operatorname{int}\Theta$ *we have*

$$\sup_{z \in \mathcal{Z}} [\ell_z(\theta_0) - \ell_z^\star(\Theta)] = \infty,$$

then $\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) = +\infty$.

The proof of this result in the general case is no more complex than in the one-dimensional case, so we simply refer to Lemma C.15.

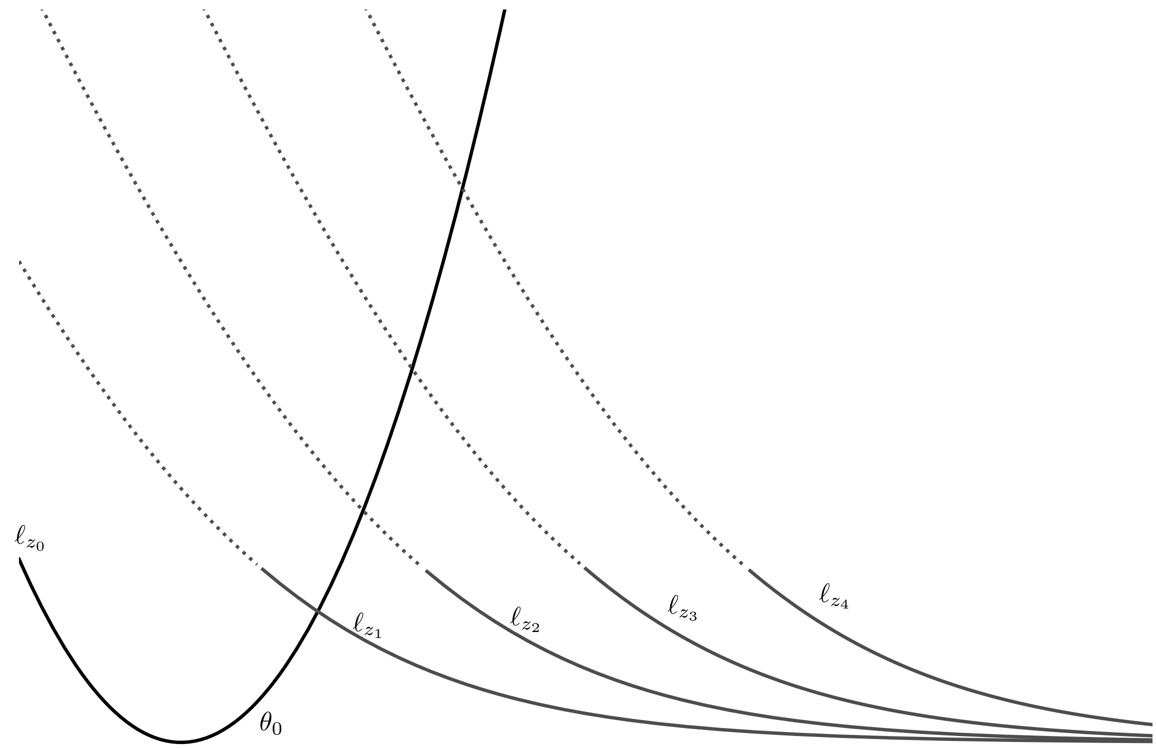

Figure 4. Choosing losses whose minima always are to the right. The loss ℓ_{z_0} has fixed positive derivative at θ_0, while $\ell_{z_1}, \ell_{z_2}, \ldots$ cannot be minimized except by taking $\theta \gg \theta_0$.

C.5.2 Planting solutions

Having shown by Lemma C.11 that unless for all $\theta \in \operatorname{int}\Theta$,

$$\sup_{z\in\mathcal{Z}}[\ell_z(\theta) - \ell_z^\star(\Theta)] < \infty, \tag{26}$$

the minimax risk is infinite, we now consider the case that Condition 2 fails while inequality (26) holds. As a consequence of inequality (26), all loss functions ℓ_z necessarily grow (or at least cannot decrease) as the parameter $|\theta| \to \infty$:

Lemma C.12. *Let inequality* (26) *hold for all* $\theta \in \operatorname{int}\Theta$. *Then*

$$\lim_{\theta\downarrow-\infty} \sup_{z\in\mathcal{Z}} \mathsf{D}_+\ell_z(\theta) \le 0 \quad \textit{and} \quad \lim_{\theta\uparrow\infty} \inf_{z\in\mathcal{Z}} \mathsf{D}_-\ell_z(\theta) \ge 0.$$

Proof We prove the left claim; the right is similar. Assume for the sake of contradiction $\sup_{z\in\mathcal{Z}} \mathsf{D}_+\ell_z(\theta) \ge a > 0$ for all $\theta \in \mathbb{R}$. Then fixing $\theta_0 \in \operatorname{int}\Theta$, let $\theta = \theta_0 - t$ for some (arbitrarily large) $t > 0$. Choose z such that $\mathsf{D}_+\ell_z(\theta) \ge a/2$. Then $\ell_z(\theta_0) \ge \ell_z(\theta) + \frac{at}{2}$, and taking $t \uparrow \infty$ would contradict that $\sup_z[\ell_z(\theta_0) - \ell_z^\star(\Theta)] < \infty$. □

We leverage this lemma and the assumption (26) to place solutions, as in our outline in Section 2.4. Assume without loss of generality that Θ is unbounded above, and as Condition 2 fails, we see that there exists $c > 0$ such that for any sequence θ_n, there exists $z_n \in \mathcal{Z}$ for which

$$\ell_{z_n}(\theta_n + n^2) - \ell_{z_n}^\star \ge c \tag{27}$$

(where we have used Lemma C.9). Figure 4 captures the basic idea: these loss functions ℓ_{z_n} "move right" so that even for $\theta_n \to \infty$, we always have $\ell_{z_n}(\theta_n) - \ell_{z_n}^\star(\Theta) \ge c$, but the minimum of the loss ℓ_{z_0} is static. We then make it so testing between the case that the loss is ℓ_{z_n} or incorporates a bit of ℓ_{z_0} is hard.

By the assumption that Θ is achievable, there exists a constant $g > 0$ and z_0 and θ_0 such that

$$\mathsf{D}_{-}\ell_{z_0}(\theta_0) \geq g.$$

Because $\theta \mapsto \mathsf{D}_{-}\ell_z(\theta)$ is non-decreasing, for all $\theta_0 \leq \theta_n \to \infty$ Lemma C.12 shows that for any $\delta > 0$, we may take N large enough that $\theta_n \geq N$ implies

$$\mathsf{D}_{-}\ell_{z_0}(\theta_n) \geq \mathsf{D}_{-}\ell_{z_0}(\theta_0) \geq g \quad \text{and} \quad \inf_{\theta \geq \theta_n} \inf_{z \in \mathcal{Z}} \mathsf{D}_{-}\ell_z(\theta) \geq -\delta,$$

while the failure of Condition 2 guarantees there exists $z_n \in \mathcal{Z}$ for which $\mathsf{D}_{-}\ell_{z_n}(\theta_n) < 0$ and inequality (27) holds.

Define the probability distributions

$$P_0 := \mathbf{1}_{z_n} \quad \text{and} \quad P_1 := \left(1 - \frac{1}{n}\right)\mathbf{1}_{z_n} + \frac{1}{n}\mathbf{1}_{z_0}.$$

By inspection (and ignoring the measure-zero sets of non-differentiable points, which are unimportant for this argument), we see that

$$L'_{P_1}(\theta) \geq -\left(1 - \frac{1}{n}\right)\delta + \frac{g}{n} \geq \frac{g}{2n} \quad \text{for} \quad \theta \geq \theta_n,$$

and so $\theta \geq \theta_n + n^2$ implies

$$L_{P_1}(\theta) = L_{P_1}(\theta_n) + \int_{\theta_n}^{\theta} L'_{P_1}(t)dt \geq L_{P_1}(\theta_n) + \int_0^{n^2} \frac{g}{2n}dt = L_{P_1}(\theta_n) + \frac{gn}{2}.$$

Similarly, $\theta \leq \theta_n + n^2$ implies $L_{P_0}(\theta) \geq L^\star_{P_0}(\Theta) + c$ by assumption (27). We therefore obtain

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) \geq \min\left\{c, \frac{gn}{2}\right\} \geq c$$

for all large n. The minimax bound (13) thus implies

$$\mathfrak{M}^{\text{low}}(\ell, \mathcal{Z}, \Theta) \geq \frac{c}{2}\left(1 - \|P_0^n - P_1^n\|_{\text{TV}}\right).$$

Applying Lemma 2.1 shows that $\|P_0^n - P_1^n\|_{\text{TV}} \leq (1 + o(1))\sqrt{1 - e^{-2}}$, giving the theorem in the one-dimensional case.

C.6 Proof of Theorem 3: the d-dimensional case

We first perform a reduction to simplify our calculations by modifying Condition 2 to a similar condition that will turn out to be equivalent, but which leverages halfspaces to more easily take advantage of convexity. To state the condition, let

$$H_{v,t} := \{\theta \mid \langle \theta, v \rangle \leq t \|v\|_2\}$$

be the closed halfspace indexed by the direction $v/\|v\|_2$ (where if $v = 0$ then obviously $H_{v,t} = \mathbb{R}^d$). Consider the following alternative version of Condition 2:

Condition 2': For all $\epsilon > 0$, there exists a $t < \infty$ such that

$$L^\star_P(H_{v,t} \cap \Theta) - L^\star_P(\Theta) \leq \epsilon$$

for each direction v and all $P \in \mathcal{P}_{\text{disc}}(\mathcal{Z})$.

Because any compact convex set Θ_0 coincides with the intersection of all closed halfspaces containing it, Condition 2' is weaker than Condition 2. Convexity, however, means that Condition 2' implies 2, making them equivalent:

Lemma C.13. *Let $L : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex, $t < \infty$, $\epsilon > 0$, and assume that*

$$L^\star(H_{v,t} \cap \Theta) < L^\star(\Theta) + \epsilon$$

for each $v \in \mathbb{R}^d$. Then

$$L^\star(t\mathbb{B}_2 \cap \Theta) \le L^\star(\Theta) + \epsilon.$$

Proof Recall the sub-optimality sets $S_\epsilon(L) := \{\theta \in \Theta \mid L(\theta) \le L^\star(\Theta) + \epsilon\}$, which are closed convex. The condition that $L^\star(H_{v,t} \cap \Theta) < L^\star(\Theta) + \epsilon$ implies that $H_{v,t} \cap \Theta \cap S_\epsilon \neq \emptyset$ for each v. The set $\Theta_t := \Theta \cap_v H_{v,t}$ is compact convex. So if $S_\epsilon \cap \Theta_t = \emptyset$, there necessarily exists a hyperplane strictly separating Θ_t from S_ϵ, meaning a vector $u \in \mathbb{S}^{d-1}$ for which

$$\inf_{\theta \in S_\epsilon} \langle u, \theta \rangle > \sup_{\theta \in \Theta_t} \langle u, \theta \rangle.$$

But we know by assumption that $H_{u,t} \cap \Theta \cap S_\epsilon \neq \emptyset$, which contradicts this inequality. □

Thus, if Condition 2' holds, then for each $\epsilon > 0$, there exists a finite t such that the compact set $\Theta_t := t\mathbb{B}_2 \cap \Theta \subset \Theta$ satisfies $\inf_{\theta \in \Theta_t}[L_P(\theta) - L_P^\star(\Theta)] \le \epsilon$ for all $P \in \mathcal{P}_{\rm disc}(\mathcal{Z})$. As a consequence, if Condition 2 fails (equivalently, 2' fails), we have the half-space analogue of the condition (27) in the one-dimensional case: there exists a $c > 0$ such that for all $t < \infty$, there is some $v \in \mathbb{S}^{d-1}$ and a $Q \in \mathcal{P}_{\rm disc}(\mathcal{Z})$ such that

$$\inf_{\theta \in H_{v,t} \cap \Theta} L_Q(\theta) \ge L_Q^\star(\Theta) + c. \tag{28}$$

This characterization will be central to our lower bounds, and for the remainder of the proof, we assume that Condition 1 holds (because if it fails, Theorem 1 gives the result). The remainder of the proof mimics the strategy we use in the one-dimensional case.

C.6.1 Eliminating infinite losses

As in Sec. C.6, we first eliminate cases in which the losses ℓ_z may tend to $-\infty$ in any way. We first eliminate the case that $\ell_z^\star(\Theta) = -\infty$ for some $z \in \mathcal{Z}$ as a triviality, as in Lemma C.10.

Lemma C.14. *If there exists $z \in \mathcal{Z}$ such that $\ell_z^\star(\Theta) = -\infty$, then*

$$\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) = +\infty.$$

Proof Let z be such that $\ell_z^\star(\Theta) = -\infty$, $\theta_0 \in \operatorname{int}\Theta$, and use the shorthand $\nabla \ell_z(\theta_0) \in \partial \ell_z(\theta_0)$. We now place gradients (recall Fig. 1) to guarantee that a particular mixture distribution has finite loss. For $v = \nabla \ell_z(\theta_0)/\left\|\nabla \ell_z(\theta_0)\right\|_2$, Lemma C.7 guarantees that there is a finitely supported distribution Q and $\beta > 0$ for which a point θ_1 of the form $\theta_1 := \theta_0 + tv + v_\perp$ for some $t \ge 0$ and $\langle v, v_\perp \rangle = 0$ satisfies

$$-\beta \nabla \ell_z(\theta_0) \in \partial L_Q(\theta_1).$$

We then obtain

$$\begin{aligned} L_Q(\theta) + \beta \ell_z(\theta) &\ge L_Q(\theta_1) + \langle -\beta \nabla \ell_z(\theta_0), \theta - \theta_1 \rangle + \beta \ell_z(\theta_0) + \beta \langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle \\ &= L_Q(\theta_1) + t\beta \left\|\nabla \ell_z(\theta_0)\right\|_2 + \beta \ell_z(\theta_0) \end{aligned}$$

by the choice of v. In particular,

$$\inf_\theta L_Q(\theta) + \beta \ell_z(\theta) \ge L_Q(\theta_1) + \beta \ell_z(\theta_0) > -\infty,$$

while $L_Q^\star(\Theta) < \infty$ and $\ell_z^\star(\Theta) = -\infty$, and for any compact subset $\Theta_0 \subset \Theta$, properness of ℓ_z guarantees $\ell_z^\star(\Theta_0) > -\infty$. Defining $p = \frac{1}{1+\beta}$, the distributions

$$P_0 = \mathbf{1}_z \quad \text{and} \quad P_1 = \frac{1}{\beta + 1} Q + \frac{\beta}{\beta + 1} \mathbf{1}_z = pQ + (1-p)\mathbf{1}_z$$

yield losses for which, evidently,

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) = +\infty.$$

As in the proof of Lemma C.10, taking $\epsilon = \frac{\beta}{\beta+1}$ yields $\|P_0 - P_1\|_{\text{TV}} \le 1 - \epsilon$, so Lemma 2.1 implies $\|P_0^n - P_1^n\|_{\text{TV}} \le \sqrt{1-\epsilon^{2n}} < 1$. Thus $\mathfrak{M}_n^{\text{low}} = +\infty$. □

Lemma C.15. *Let Condition 1 hold. If for some $\theta_0 \in \text{int}\,\Theta$ we have*

$$\sup_{z \in \mathcal{Z}} [\ell_z(\theta_0) - \ell_z^\star(\Theta)] = \infty,$$

then $\mathfrak{M}_n^{\text{low}}(\ell, \mathcal{Z}, \Theta) = +\infty$.

Proof Lemma C.14 allows us to assume that $\ell_z^\star(\Theta) > -\infty$ for each $z \in \mathcal{Z}$ (otherwise, $\mathfrak{M}_n^{\text{low}} = +\infty$ in any case). By Condition 1, there exists $M_0 < \infty$ such that $\partial \ell_z(\theta_0) \subset M_0 \mathbb{B}_2$ for each $z \in \mathcal{Z}$. Lemma C.7 thus guarantees that there exist constants $\alpha_0, \beta_0 > 0$ such that for any $z \in \mathcal{Z}$ and any $\nabla \ell_z(\theta_0) \in \partial \ell_z(\theta_0)$, there is a finitely supported distribution Q and point of the form $\theta_1 = \theta_0 + tv + v_\perp$ for the direction $v = \nabla \ell_z(\theta_0)/\left\|\nabla \ell_z(\theta_0)\right\|_2$ and some $\alpha_0 \le t \le \frac{1}{\alpha_0}$ for which

$$-\beta \nabla \ell_z(\theta_0) \in \partial L_Q(\theta_1) \quad \text{for a } \beta \ge \beta_0.$$

(In Figure 1, this corresponds to the choice $\theta_1 = \theta_0 + tv + v_\perp$ for $v = \nabla \ell_z(\theta_0)/\left\|\nabla \ell_z(\theta_0)\right\|_2$, though in this case $\nabla \ell_z(\theta_0)$ is bounded for all z, while the value $\ell_z(\theta_0) - \ell_z^\star$ may be arbitrary.)

Choose $p = \frac{1}{1+\beta/2}$, so that $\beta p - (1-p) = \frac{\beta}{2+\beta}$, and so

$$\begin{aligned}
& pL_Q(\theta) + (1-p)\ell_z(\theta) \\
& \stackrel{(i)}{\ge} p\left[L_Q(\theta_1) + \langle -\beta \nabla \ell_z(\theta_0), \theta - \theta_1 \rangle\right] + (1-p)\left[\ell_z(\theta_0) + \langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle\right] \\
& \stackrel{(ii)}{=} pL_Q(\theta_1) + (1-p)\ell_z(\theta_0) + p\langle -\beta \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle + tp\beta \left\|\nabla \ell_z(\theta_0)\right\|_2 + (1-p)\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle \\
& = pL_Q(\theta_1) + (1-p)\ell_z(\theta_0) + tp\beta \left\|\nabla \ell_z(\theta_0)\right\|_2 - \frac{\beta}{2+\beta}\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle.
\end{aligned}$$

Here, inequality (i) follows from the first-order convexity condition, while inequality (ii) uses the definition of θ_1.

Define the two point distributions

$$P_0 = \mathbf{1}_z \quad \text{and} \quad P_1 = (1-p)\mathbf{1}_z + pQ.$$

Letting $\gamma = \frac{\beta}{2+\beta}$ for shorthand, noting that $\gamma \ge \frac{\beta_0}{2+\beta_0}$ regardless of the choice of z, we find that for any $\nabla L_Q(\theta_0) \in \partial L_Q(\theta_0)$, the first order condition for convexity implies

$$\begin{aligned}
L_{P_1}(\theta) &\ge pL_Q(\theta_1) + (1-p)\ell_z(\theta_0) - \gamma\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle \\
&\ge L_{P_1}(\theta_0) + p\langle \nabla L_Q(\theta_0), \theta_1 - \theta_0 \rangle - \gamma\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle \\
&\ge L_{P_1}(\theta_0) - \frac{2pM_0}{\alpha_0} - \gamma\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle,
\end{aligned}$$

where the final inequality follows via Cauchy-Schwarz.

We now demonstrate the separation in optimization distance (12). By the preceding display, for any $c > 0$

$$L_{P_1}(\theta) \le L_{P_1}(\theta_0) + c \ \text{ implies } \ -\gamma\langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle \le c + \frac{2M_0}{\alpha_0}.$$

Because the choice of $z \in \mathcal{Z}$ was arbitrary, we may assume that for any $K < \infty$, we choose z so that $\ell_z(\theta_0) - \ell_z^\star(\Theta) \ge K$. Then

$$\ell_z(\theta) \le \ell_z^\star + K/2 \ \text{ implies } \ \ell_z(\theta) \le \ell_z(\theta_0) - K/2 \ \text{ so } \ -\frac{K}{2} \ge \langle \nabla \ell_z(\theta_0), \theta - \theta_0 \rangle.$$

Setting $c = \frac{\gamma K}{2} - \frac{2M_0}{\alpha_0}$ then yields that for *any* large enough (but finite) K, we can choose distributions P_0 and P_1 on $\mathcal{Z}$ for which

$$d_{\mathsf{opt}}(L_{P_1}, L_{P_0}) \ge \min\{\gamma, 1\} \cdot \frac{K}{2} \ge \min\left\{\frac{\beta_0}{2+\beta_0}, 1\right\} \cdot \frac{K}{2}$$

while

$$\|P_0 - P_1\|_{\rm TV} \le p = \frac{1}{1+\beta/2} \le \frac{2}{2+\beta_0}.$$

Applying Lemma 2.1, $\|P_0^n - P_1^n\|_{\rm TV} \le \sqrt{1-(1-p)^{2n}} < 1$. Take K arbitrarily large. $\square$

C.6.2 Planting solutions

We now turn to analogies of the arguments in Section C.5.2 that allow us to show separation in population losses by planting solutions appropriately. The first step in this development is to show that, so long as $\inf_z \ell_z^\star(\Theta) > -\infty$, all loss functions must eventually grow away from any $\theta_0 \in \operatorname{int}\Theta$, as in Lemma C.12. To state the result, recall the left derivative $\mathsf{D}_- f(\theta; v) \coloneqq \inf\{\langle g, v\rangle \mid g \in \partial f(\theta)\}$ from Section C.1.2.

Lemma C.16. *Let Condition 1 hold and assume* $\lim_n \mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) < \infty$*. Then for any compact set* $\Theta_0 \subset \operatorname{int}\Theta$,

$$\liminf_{t\to\infty} \inf_{\|\theta\|_2 \ge t, \theta\in\Theta} \inf_{z\in\mathcal{Z}} \inf_{\theta_0\in\Theta_0} \mathsf{D}_-\ell_z\left(\theta; \frac{\theta-\theta_0}{\|\theta-\theta_0\|_2}\right) \ge 0.$$

Proof Because it is the supremum of closed convex functions, the function

$$\overline{\ell}(\theta) \coloneqq \sup_{z\in\mathcal{Z}}[\ell_z(\theta) - \ell_z(\Theta)^\star]$$

is closed convex. By Lemma C.15, $\overline{\ell}(\theta_0) < \infty$ for any $\theta_0 \in \operatorname{int}\Theta$, and as any convex function is continuous on the interior of its domain (Hiriart-Urruty and Lemaréchal, 1993a, Ch. V), the function $\overline{\ell}$ is evidently continuous on Θ_0. It therefore attains its infimum and supremum on Θ_0, whence $\sup_{\theta_0\in K} \overline{\ell}(\theta_0) < \infty$.

For the sake of contradiction, assume that the limit infimum in the lemma statement is $-2c$, where $c > 0$. Then there is some finite $t > 0$, $\theta \in \Theta$ satisfying $\operatorname{dist}(\theta, \Theta_0) \ge t$, and $\theta_0 \in \Theta_0$ and $z \in \mathcal{Z}$ for which

$$\mathsf{D}_-\ell_z(\theta; v) \le -c,$$

where we defined $v = (\theta - \theta_0)/\|\theta-\theta_0\|_2$ for shorthand. The monotonicity of (sub)gradients then implies

$$\mathsf{D}_-\ell_z\left(u\theta + (1-u)\theta_0; v\right) \le \mathsf{D}_-\ell_z(\theta; v) \le -c \quad \text{for } u \in [0,1].$$

For any convex f and $\theta_0 \in \operatorname{int}\operatorname{dom} f$, we have the integral form (Hiriart-Urruty and Lemaréchal, 1993a, Thm. VI.2.3.4)

$$f(\theta) = f(\theta_0) + \int_0^1 \mathsf{D}f(\theta_0 + t(\theta-\theta_0); \theta-\theta_0)dt = f(\theta_0) + \int_0^1 \mathsf{D}_- f(\theta_0 + t(\theta-\theta_0); \theta-\theta_0)dt.$$

Applying this to ℓ_z, we see $\ell_z(\theta) = \ell_z(\theta_0) + \int_0^1 \mathsf{D}_-\ell_z(\theta_0 + u(\theta-\theta_0); v)du \cdot \|\theta_0-\theta\|_2 \le \ell_z(\theta_0) - ct$. For t large enough, this contradicts that $\sup_{\theta_0\in\Theta_0}\sup_z[\ell_z(\theta_0) - \ell_z^\star] < \infty$. $\square$

Because integration and directional differentiation commute for closed convex functions Bertsekas (1973), we may replace the infimum over $z \in \mathcal{Z}$ with an infimum over $P \in \mathcal{P}_{\rm disc}(\mathcal{Z})$, so that

$$\liminf_{t\to\infty} \inf_{\|\theta\|_2 \ge t, \theta\in\Theta} \inf_{P\in\mathcal{P}_{\rm disc}(\mathcal{Z})} \inf_{\theta_0\in\Theta_0} \mathsf{D}_- L_P\left(\theta; \frac{\theta-\theta_0}{\|\theta-\theta_0\|_2}\right) \ge 0 \tag{29}$$

for any compact $\Theta_0 \subset \operatorname{int}\Theta$.

Now we provide an analogue of Lemma C.7 that allows us to plant solutions more carefully, constructing a finitely supported distribution $\overline{Q}$ such that $L_{\overline{Q}}$ is coercive and its minimum belongs to a compact set. This construction mimics the distinct point θ_0 in the one-dimensional case (Fig. 4 and ℓ_{z_0} there), but requires more care.

Lemma C.17. *Let $\theta_0 \in \operatorname{int}\Theta$. There exists a finitely supported distribution $\overline{Q}$ and values $\gamma > 0$ and $b < \infty$ such that*

$$\mathsf{D}L_{\overline{Q}}\left(\theta; \frac{\theta - \theta_0}{\|\theta - \theta_0\|_2}\right) \geq \gamma$$

whenever $\theta \in \Theta$, $\|\theta - \theta_0\|_2 \geq b$.

Proof Let $\mathcal{E} = \{e_i, -e_i\}_{i=1}^d$ be the collection of standard basis vectors and their negations. Use Lemma C.7 to find a convex compact set $K \subset \Theta$ such that for each $u \in \mathcal{E}$, there exsists a finitely supported distribution Q_u and scalar $\alpha_u > 0$ such that

$$\alpha_u u \in \partial L_{Q_u}(\theta_u) \text{ for some } \theta_u \in K.$$

Thus for each $u \in \mathcal{E}$,

$$L_{Q_u}(\theta) - L_{Q_u}(\theta_u) \geq \alpha_u \langle \theta - \theta_u, u \rangle$$

By Lemma C.14, we see that w.l.o.g. we may assume $L^\star_{Q_u}(\Theta) = 0$, and so we also have $L_{Q_u}(\theta) - L_{Q_u}(\theta_u) \geq -L_{Q_u}(\theta_u)$. We obtain

$$\frac{1}{2d}\sum_{u\in\mathcal{E}}(L_{Q_u}(\theta) - L_{Q_u}(\theta_u)) \geq \frac{1}{2d}\sum_{u\in\mathcal{E}}\left\{\alpha_u\langle \theta - \theta_u, u\rangle \vee -L_{Q_u}(\theta_u)\right\}.$$

Letting $\alpha_\star = \min_{u\in\mathcal{E}} \alpha_u > 0$, we see that once $\alpha_\star \min_{u\in\mathcal{E}} \|\theta - \theta_u\|_\infty \geq 2d \max_{u\in\mathcal{E}} L_{Q_u}(\theta_u)$,

$$\frac{1}{2d}\sum_{u\in\mathcal{E}}(L_{Q_u}(\theta) - L_{Q_u}(\theta_u)) \geq \frac{1}{2d}\alpha_\star \inf_{\theta_1\in K}\|\theta - \theta_1\|_\infty - \max_{u\in\mathcal{E}} L_{Q_u}(\theta_u) > 0.$$

In particular, $L_{\overline{Q}}$ grows at least linearly outside of a neighborhood of K for $\overline{Q} = \frac{1}{2d}\sum_{u\in\mathcal{U}} Q_u$. As $\operatorname{dist}(\theta_0, K) < \infty$, this linear growth also applies far from θ_0, which implies the lemma. ☐

C.6.3 The optimization separation and testing lower bound

With Lemma C.17 in hand, we can now construct hard instances extending the construction in the one-dimensional case (Fig. 4) to the full-dimensional setting. Let $\overline{Q}$ be the finitely supported distribution Lemma C.17 promises with the attendant growth constant $\gamma > 0$, and let $0 < \delta \leq 1$ be a value to be chosen. Fix $\theta_0 \in \operatorname{int}\Theta$, and for $\theta \in \mathbb{R}^d$, define the direction $v_0(\theta) := (\theta - \theta_0)/\|\theta - \theta_0\|_2$. Because Condition 2' fails, we may take $t < \infty$ large enough that it satisfies the following four conditions:

(i) There exists v with $\|v\|_2 = 1$ and $Q \in \mathcal{P}_{\text{disc}}(\mathcal{Z})$ for which

$$L^\star_Q(H_{v,t} \cap \Theta) \geq L^\star_Q(\Theta) + c.$$

(ii) For any $\theta \in \Theta$ satisfying $\langle v, \theta\rangle \geq \frac{t}{2}$,

$$\mathsf{D}L_{\overline{Q}}(\theta; v_0(\theta)) \geq \gamma.$$

(iii) For any $\theta \in \Theta$ satisfying $\langle v, \theta\rangle \geq \frac{t}{2}$,

$$\mathsf{D}L_P(\theta; v_0(\theta)) \geq -\frac{\delta\gamma}{2n} \text{ for all } P \in \mathcal{P}_{\text{disc}}(\mathcal{Z}).$$

(iv) $t \geq \frac{4cn}{\gamma\delta}$.

The conditions may appear to be circular, as parts (ii) and (iii) repose on the direction v in part (i), but this is not truly an issue. We may satisfy the requirement (ii) via Lemma C.17, so long as t is large, because $\|\theta\|_2 \geq \frac{t}{2}$ certainly implies $\langle v, \theta\rangle \geq \frac{t}{2}$. To satisfy the requirement (iii), we similarly use Lemma C.16 (actually, the remark (29) following the lemma). Satisfying requirement (iv) is trivial; we may always simply take t larger in the other parts, and the failure of Condition 2' is sufficient to guarantee (i).

With these four requirements satisfied, consider the two distributions

$$P_0 \coloneqq Q \;\;\text{and}\;\; P_1 \coloneqq \left(1 - \frac{\delta}{n}\right) Q + \frac{\delta}{n}\overline{Q}.$$

Let θ_1 be any vector satisfying $\langle v, \theta_1 \rangle \ge t$, and let $\theta_{1/2}$ be the point at which the line segment $[u\theta_1 + (1-u)\theta_0]$ crosses the hyperplane $\langle v, \theta \rangle = t/2$, that is, at $u = \frac{t}{2\langle v, \theta_1 - \theta_0 \rangle}$. Then $u < \frac{1}{2}$, and the directions $v_0(\theta_1) = v_0(\theta_{1/2}) = v_0(u\theta_{1/2} + (1-u)\theta_1)$ for all $u \in [0,1]$. Using the shorthand $\theta_u = u\theta_{1/2} + (1-u)\theta_1$ and computing directional derivatives, we see that $u \in [0,1]$ implies

$$\mathsf{D}L_{P_1}(\theta_u; v_0(\theta_1)) = \left(1 - \frac{\delta}{n}\right) \mathsf{D}L_Q(\theta_u; v_0(\theta_1)) + \frac{\delta}{n}\mathsf{D}L_{\overline{Q}}(\theta_u; v_0(\theta_1)) \ge \frac{\delta\gamma}{n} - \left(1 - \frac{\delta}{n}\right)\frac{\delta\gamma}{2n},$$

where the inequality follows from the assumptions (ii) and (iii) on t. Thus

$$\begin{aligned} L_{P_1}(\theta_1) &= L_{P_1}(\theta_{1/2}) + \left\|\theta_1 - \theta_{1/2}\right\|_2 \int_0^1 \mathsf{D}L_{P_1}\left((1-u)\theta_{1/2} + u\theta_1; v_0(\theta_1)\right) du \\ &\ge L_{P_1}(\theta_{1/2}) + \left\|\theta_1 - \theta_{1/2}\right\|_2 \frac{\delta\gamma}{2n} \\ &\ge L_{P_1}(\theta_{1/2}) + \frac{t}{2} \cdot \frac{\delta\gamma}{2n} \ge L_{P_1}(\theta_{1/2}) + c, \end{aligned}$$

where the final step uses the choice (iv) that $t \ge \frac{4cn}{\delta\gamma}$.

Performing a similar calculation, we see that if $\langle v, \theta_1 \rangle \le t$, then

$$L_{P_0}(\theta) = L_Q(\theta) \ge L_Q^\star(\Theta) + c = L_{P_0}^\star(\Theta) + c$$

by assumption (i). Taking the contrapositive, we see that if $L_{P_1}(\theta) < c$, then necessarily $\langle \theta, v \rangle < t$, implying that $L_{P_0}(\theta) \ge L_{P_0}^\star(\Theta) + c$, while if $L_{P_0}(\theta) < L_{P_0}^\star(\Theta) + c$, then $\langle \theta, v \rangle \ge t$ and $L_{P_1}(\theta) \ge L_{P_1}^\star(\Theta) + c$. That is,

$$d_{\mathsf{opt}}(L_{P_0}, L_{P_1}) \ge c.$$

Computing the variation distance between the probability distributions is straightforward: we observe that

$$\|P_0^n - P_1^n\|_{\rm TV} \le \sqrt{1 - (1 - \delta/n)^n}\sqrt{1 + (1 - \delta/n)^n} \to \sqrt{1 - e^{-2\delta}}.$$

Thus, for any $\delta > 0$ and large enough n, we may choose P_0 and P_1 so that

$$\mathfrak{M}_n^{\rm low}(\ell, \mathcal{Z}, \Theta) \ge \frac{d_{\mathsf{opt}}(L_{P_0}, L_{P_1})}{2}\left(1 - \sqrt{1 - e^{-2\delta}}\right) \ge \frac{c}{2}\left(1 - \sqrt{1 - e^{-2\delta}}\right).$$

Take $\delta \downarrow 0$ and recognize that $\mathfrak{M}_n^{\rm low}$ is non-increasing in n.

C.7 Proof of Proposition 2

As a first step, we must show that Condition 2 extends so that inequality (10) actually holds for all $P \in \mathcal{P}_\ell(\mathcal{Z})$, where $\mathcal{P}_\ell(\mathcal{Z})$ denotes the well-defined probabilities (9).

Lemma C.18. *Let Condition 2 hold. Then for all $\epsilon > 0$, there is a compact $\Theta_0 \subset \Theta$ such that*

$$\sup_{Q \in \mathcal{P}_\ell(\mathcal{Z})} L_Q^\star(\Theta_0) - L_Q^\star(\Theta) \le \epsilon. \tag{30}$$

See Appendix B.4 for a proof of this result.

Now let $\epsilon > 0$ and $\Theta_0 \subset \Theta$ be a compact convex set for which $\inf_{\theta \in \Theta_0}[L_Q(\theta) - L_Q^\star(\Theta)] \le \epsilon$ for all $Q \in \mathcal{P}_\ell(\mathcal{Z})$, as Lemma C.18 promises. Then Theorem 2 shows that there exists an estimator $\widehat{\theta}$ taking values in Θ_0 and some $N < \infty$ such that $n \ge N$ implies

$$\sup_{P \in \mathcal{P}_\ell} \mathbb{E}_{P^n}\left[L_P(\widehat{\theta}(Z_1^n)) - L_P^\star(\Theta_0)\right] \le \epsilon.$$

Of course, $L_P^\star(\Theta) \le L_P^\star(\Theta_0) + \epsilon$, so $\mathfrak{M}_n(\ell, \mathcal{Z}, \Theta) \le 2\epsilon$.

D Proofs of results related to stationary points

D.1 Proof of Lemma 3.1

Part (i) is immediate. For claim (ii), note that $\partial L(\theta) = [a, b]$, where a and b satisfy either $-\infty < a \le b < \infty$, $a = -\infty$, or $b = +\infty$. Considering each of the cases, it becomes apparent that we need only consider the case that $a > 0$ (as the $b < 0$ case is similar), so w.l.o.g. we assume $a > 0$, yielding $\mathsf{D}_- L(\theta) = \inf \partial L(\theta) > 0$, so $\epsilon_L(\theta) > 0$.

For the final claim, note that if $\operatorname{argmin} L = \emptyset$, then necessarily L is strictly monotone (as otherwise, it would have a stationary point and hence a minimizer), yielding the one-sidedness of $\mathsf{D}_+ L$ or $\mathsf{D}_- L$. Because $\mathsf{D}_+ L$ and $\mathsf{D}_- L$ are monotone non-decreasing, then considering the case that $\mathsf{D}_+ L < 0$, the ordering in part (iii) of Lemma C.1 gives that $\mathsf{D}_- L < 0$ as well, and

$$\lim_{\theta \to \infty} \mathsf{D}_+ L(\theta) = \lim_{\theta \to \infty} \mathsf{D}_- L(\theta) = \sup_{\theta} \mathsf{D}_+ L(\theta) = \sup_{g, \theta} \{g \in \partial L(\theta)\} \le 0.$$

That is, $\lim_{\theta \to \infty} \epsilon_L(\theta) = \lim_{\theta \to \infty} -\mathsf{D}_+ L(\theta) - \lim_{\theta \to \infty} (-\mathsf{D}_+ L(\theta)) = 0$.

D.2 Proof of Proposition 3

Let $L = L_P$ for shorthand when convenient. We first give the argument assuming that $\operatorname{argmin} L = [\theta_0, \theta_1]$ is non-empty, then extend it to the case that it is empty. Fixing $t > 0$, consider the event that $\epsilon_L(\widehat{\theta}) > t$. For simplicity and without loss of generality, let us assume that $\mathsf{D}_- L(\widehat{\theta}) > t$. Define

$$\theta_t \coloneqq \inf \{\theta \mid \mathsf{D}_- L(\theta) \ge t\} = \sup \{\theta \mid \mathsf{D}_- L(\theta) < t\},$$

where the equality follows from Lemma C.1, part (iii), as $\theta < \theta_t$ implies $\mathsf{D}_- L(\theta) < t$ and $\mathsf{D}_- L(\theta_t) = \lim_{\theta \uparrow \theta_t} \mathsf{D}_- L(\theta) = \lim_{\theta \uparrow \theta_t} \mathsf{D}_+ L(\theta) \le t$. We also observe that $\mathsf{D}_+ L(\theta_t) = \lim_{\theta \downarrow \theta_t} \mathsf{D}_+ L(\theta) \ge t$, again by Lemma C.1.

Combining the observation that $\mathsf{D}_- L(\widehat{\theta}) > t$ and $\mathsf{D}_- L(\theta_t) \le t$ imply $\widehat{\theta} > \theta_t$ with the preceding derivations, we see that $\widehat{\theta} \in \operatorname{argmin} L_{P_n}$ implies $\mathsf{D}_+ L_{P_n}(\theta_t) \le \mathsf{D}_- L_{P_n}(\widehat{\theta}) \le 0$, so

$$\begin{aligned}
\mathbb{P}\left(\mathsf{D}_- L(\widehat{\theta}) > t\right) &\le \mathbb{P}\left(\mathsf{D}_+ L_{P_n}(\theta_t) \le 0\right) = \mathbb{P}\left(\mathsf{D}_+ L_{P_n}(\theta_t) - \mathsf{D}_+ L_P(\theta_t) \le -\mathsf{D}_+ L_P(\theta_t)\right) \\
&\le \mathbb{P}\left(\mathsf{D}_+ L_{P_n}(\theta_t) - \mathsf{D}_+ L_P(\theta_t) \le -t\right) \le \exp\left(-\frac{nt^2}{2M^2}\right),
\end{aligned}$$

because $\mathsf{D}_+ L_{P_n}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathsf{D}_+ \ell_{Z_i}(\theta)$ (cf. Bertsekas (1973)), where we have applied a Hoeffding bound. A completely parallel calculation gives that

$$\mathbb{P}\left(\mathsf{D}_+ L(\widehat{\theta}) < -t\right) \le \exp\left(-\frac{nt^2}{2M^2}\right)$$

for any $t > 0$.

Lastly, we address the case that $\operatorname{argmin} L = \emptyset$. Without loss of generality, assume that $\mathsf{D}_+ L > 0$, so that L is increasing, and $\mathsf{D}_- L(-\infty) = \lim_{\theta \to -\infty} \mathsf{D}_- L(\theta)$ satisfies $\mathsf{D}_- L(-\infty) = \inf_{\theta, g} \{g \in \partial L(\theta)\} \ge 0$ and $\epsilon_L(\theta) = \mathsf{D}_- L(\theta) - \mathsf{D}_- L(-\infty)$. If $\widehat{\theta} = -\infty$, then $\epsilon_L(\widehat{\theta}) = 0$, so we assume now that $\widehat{\theta}$ is finite. Consider the event that $\epsilon_L(\widehat{\theta}) > t$, i.e., $\mathsf{D}_- L(\widehat{\theta}) > \mathsf{D}_- L(-\infty) + t$. As above, defining

$$\theta_t = \inf \{\theta \mid \mathsf{D}_- L(\theta) - \mathsf{D}_- L(-\infty) \ge t\},$$

we have $\mathsf{D}_+ L(\theta_t) \ge t + \mathsf{D}_- L(-\infty) \ge t$, and $\widehat{\theta} > \theta_t$. Then $\mathsf{D}_- L_{P_n}(\theta_t) \le \mathsf{D}_- L_{P_n}(\widehat{\theta}) \le 0$, while

$$\begin{aligned}
\mathbb{P}\left(\mathsf{D}_- L(\widehat{\theta}) > \mathsf{D}_- L(-\infty) + t\right) &\le \mathbb{P}\left(\mathsf{D}_+ L_{P_n}(\theta_t) \le 0\right) \\
&= \mathbb{P}\left(\mathsf{D}_+ L_{P_n}(\theta_t) - \mathsf{D}_+ L_P(\theta_t) \le -\mathsf{D}_+ L_P(\theta_t)\right) \le \exp\left(-\frac{nt^2}{2M^2}\right).
\end{aligned}$$