JMLR: Workshop and Conference Proceedings vol 40:1–54, 2015

Convex Risk Minimization and Conditional Probability Estimation

Matus Telgarsky MTELGARS@CS.UCSD.EDU
University of Michigan, Ann Arbor

Miroslav Dudík MDUDIK@MICROSOFT.COM
Microsoft Research

Robert Schapire SCHAPIRE@MICROSOFT.COM
Microsoft Research and Princeton University

Abstract

This paper proves, in very general settings, that convex risk minimization is a procedure to select a unique conditional probability model determined by the classification problem. Unlike most previous work, we give results that are general enough to include cases in which no minimum exists, as occurs typically, for instance, with standard boosting algorithms. Concretely, we first show that any sequence of predictors minimizing convex risk over the source distribution will converge to this unique model when the class of predictors is linear (but potentially of infinite dimension). Secondly, we show the same result holds for *empirical* risk minimization whenever this class of predictors is finite dimensional, where the essential technical contribution is a norm-free generalization bound.

Keywords: Convex duality, classification, conditional probability estimation, maximum entropy, consistency, Orlicz spaces.

1. Introduction

The goal in (binary) classification is to learn to accurately predict the label $y \in \{-1, +1\}$ associated with an input x. Unfortunately, it is NP-hard even to approximate this problem in easy cases (Guruswami and Raghavendra, 2006); thus a computationally attractive surrogate is often utilized. Foremost amongst these is *convex risk minimization* in which a sequence of predictors are produced which minimize in the limit some convex upper bound on a predictor's classification error. In this paper, we attempt to analyze the effectiveness of such methods in as much generality as possible. Specifically, we aim to address the following questions:

(Q1) Suppose a sequence of predictors minimizes the convex risk over the true distribution. Does this sequence converge to some concrete object? This question is murky because convex functions need not have a minimum; for instance, the function e^x has no minimum, but rather is minimized in the limit $x \to -\infty$. For the high-dimensional problems considered in convex risk minimization, the minimum may also only occur "at infinity" but in a far less straightforward way. This is typically the case, for instance, for standard boosting algorithms like AdaBoost (Schapire and Freund, 2012). In such cases, what can be said concretely about the convergence of a minimizing sequence?

(Q2) Now suppose a given sequence of predictors minimizes the *empirical* convex risk, meaning the convex risk over some finite random draw from the distribution. What can be said about

convergence with respect to the true distribution? In other words, what can be said about generalization and learning? The resolution is unclear here as well, since the preceding question highlights the need for predictors to be arbitrarily large, thus dooming the boundedness on which most standard statistical procedures rely (Boucheron et al., 2005, Section 4).

In this paper we resolve both these questions by showing that convex risk minimization converges to a unique conditional probability model $\bar{\eta}$.

Main results. To state our main theorems, we first present our learning setting. We consider *linear* classes of functions. That is, given a base set $\mathcal{H}$ of prediction functions $h : \mathcal{X} \to [-1, +1]$, the corresponding linear class consists of weightings of these functions described by weight vectors w with $\sum_{h\in\mathcal{H}}|w[h]| < \infty$ where $w[h]$ denotes the weight of the function h, and where it is understood that these weights are non-zero only on a countable subset of $\mathcal{H}$. Formally, this class is

$$\Big\{x \mapsto \sum_{h\in\mathcal{H}} w[h]\, h(x) : \sum_{h\in\mathcal{H}} \big|w[h]\big| < \infty\Big\}.$$

This setting recovers, for instance, the classical regression setting by choosing $\mathcal{H}$ to consist of covariates corresponding to the dimensions of x, as well as the classical boosting setting by leaving $\mathcal{H}$ arbitrary. Let $L_1(\mathcal{H})$ denote all possible choices for w as above; moreover, given $w \in L_1(\mathcal{H})$, let $Hw : \mathcal{X} \to \mathbb{R}$ denote the corresponding element of the linear class, meaning, $(Hw)(x) = \sum_h w[h]\, h(x)$. Thus, H is a linear operator, abstractly collecting the elements of $\mathcal{H}$ as "columns".

The loss functions ℓ that we study come from a large class $\mathbb{L}_{\mathrm{b}}^{2+}$ of certain twice continuously differentiable losses, whose precise definition is deferred to Section 3. Both the well-studied logistic loss $\ell_{\log}(r) := \ln(1 + \exp(r))$ and exponential loss $\ell_{\exp}(r) := \exp(r)$ belong to $\mathbb{L}_{\mathrm{b}}^{2+}$. With respect to loss ℓ, we define the population and empirical convex risk to be

$$\mathcal{R}(w) := \int \ell\Big(-y(Hw)(x)\Big)\, d\mu(x, y) \qquad \text{and} \qquad \hat{\mathcal{R}}_n(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\Big(-y_i(Hw)(x_i)\Big),$$

where $((x_i, y_i))_{i=1}^n$ is an i.i.d. draw of size n from the true distribution μ. Lastly, we define the excess convex risk $\mathcal{E}(w) := \mathcal{R}(w) - \inf_{v\in L_1(\mathcal{H})} \mathcal{R}(v)$, with $\hat{\mathcal{E}}_n$ defined analogously.

There are well-established methods for converting the models produced using convex risk minimization into *conditional probability models*. Specifically, given loss $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$, functions $\mathcal{H}$, and weighting $w \in L_1(\mathcal{H})$, we define

$$\phi(r) := \frac{\ell'(r)}{\ell'(r) + \ell'(-r)} \qquad \text{and} \qquad \eta_w(x, y) := \phi\Big(y(Hw)(x)\Big). \tag{1}$$

This function $\eta_w(x, y)$, which is well-defined with range $[0, 1]$ for all $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$, can be regarded as an estimate of the conditional probability $\Pr[Y = y|x]$ (Friedman et al., 2000; Zhang, 2004; Bartlett et al., 2006). For example, logistic loss $\ell_{\log}$ yields the usual sigmoid $\phi(r) = (1 + \exp(-r))^{-1}$.

Our convergence results do not apply to the weighting sequences $(w_i)_{i\geq 1}$ directly, since, as earlier mentioned, these will often have no limit. Instead we prove convergence of their corresponding conditional probability models. Specifically, our first main result, the resolution of **(Q1)**, states that minimizing $\mathcal{R}$ implies convergence to a unique conditional probability model $\bar{\eta}$.

Theorem 1.1 *Let loss $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$, probability measure μ, and hypotheses $\mathcal{H}$ be given. Then there exists a unique conditional probability model $\bar{\eta} : \mathcal{X} \times \{-1, +1\} \to [0, 1]$ and a function $f_1 : \mathbb{R} \to \mathbb{R}_+$ with $f_1(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$ such that every $w \in L_1(\mathcal{H})$ satisfies*

$$\int |\bar{\eta}(x, y) - \eta_w(x, y)| \, d\mu(x, y) \leq f_1 \left(\mathcal{E}(w) \right).$$

In particular, every sequence $(w_i)_{i \geq 1}$ with $\lim_{i \to \infty} \mathcal{E}(w_i) = 0$ satisfies $\eta_{w_i} \to \bar{\eta}$ in $L_1(\mu)$.

Note that the existence of $\bar{\eta}$ is not immediate given the existence of sequences minimizing $\mathcal{R}$ since the collection of mappings from $\mathcal{X}$ to $[0, 1]$ is not compact in the $L_1(\mu)$ metric in general. Instead, the proof here constructs $\bar{\eta}$ directly via duality, and thereafter uses duality to control these sequences.

Theorem 1.1 carries two essential consequences. First, our analysis provides a convergence concept for algorithms utilizing convex risk minimization that is more general than previous approaches in the sense that it can handle, for instance, the unregularized boosting methods of Zhang and Yu (2005, Algorithm 1), or even any regularized scheme with regularization weakening to zero. Secondly, the real-valued model Hw can be used for classification simply by taking its sign, which is exactly equivalent to the sign of $\eta_w(\cdot, 1) - 1/2$, that is, the more likely label according to the corresponding conditional probability model η_w. Therefore, convergence properties of $(\eta_{w_i})_{i \geq 1}$ imply convergence properties of the classification errors made by $(Hw_i)_{i \geq 1}$, complementary to the results of Bartlett et al. (2006) and Zhang (2004); see Proposition 1.3.

Next comes the resolution of **(Q2)**: under the assumption $|\mathcal{H}| < \infty$, we show that it suffices to minimize the empirical risk $\hat{\mathcal{R}}_n$.

Theorem 1.2 *Suppose the setting of Theorem 1.1, in particular the existence of $\bar{\eta}$, but additionally that $|\mathcal{H}| < \infty$. There exists a nonincreasing function $f_2 : \mathbb{R} \to \mathbb{R}_+$ such that, with probability at least $1 - \delta$ over an i.i.d. draw of size $n \geq \Omega(\ln(1/\delta))$ from μ, every $w \in L_1(\mathcal{H})$ satisfies*

$$\int |\bar{\eta}(x, y) - \eta_w(x, y)| \, d\mu(x, y) = \mathcal{O}\left(f_2\left(\hat{\mathcal{E}}_n(w) \right) \left(\sqrt{\hat{\mathcal{E}}_n(w)} + \sqrt{\frac{\ln(n) + \ln(1/\delta)}{n}} \right) \right),$$

where $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ omit constants based on $\mathcal{H}$, ℓ, and μ, but not on the sample, or on w. In particular, any sequence $(w_i)_{i \geq 1}$ with $\lim_{i \to \infty} \hat{\mathcal{E}}_i(w_i) = 0$ satisfies $\eta_{w_i} \to \bar{\eta}$ in $L_1(\mu)$ a.s.

Note that perhaps the most natural approach to proving this theorem—namely, to apply properties of Rademacher complexity of Lipschitz functions (Boucheron et al., 2005)—introduces a dependence on the norm of $\|w\|$. Instead, the bound above only exhibits a dependence on $\hat{\mathcal{E}}_n(w)$, which can be made arbitrarily small by considering only nearly optimal choices. Depending on $\hat{\mathcal{E}}_n(w)$ rather than $\|w\|$ is essential as these minimizing sequences will generally exhibit unboundedly growing norms, a fact often encountered in practice (see Appendix D). Note that while Theorem 1.2 requires strictly convex losses, it is proved via generalization bounds which can handle more than just $\mathbb{L}_{\mathrm{b}}^{2+}$, in particular the hinge loss (see Lemmas 3.10 and J.9).

Illustrative example. Suppose $\mathcal{X} = [-1, 1]^2$ and $\mathcal{H}$ consists of the coordinate functions. Consider $\ell = \ell_{\log}$, i.e., logistic regression. Suppose that the measure μ puts all of the mass on points x that fall into two well-separated rectangular regions (depicted as red and blue in Figure 1), with points in the blue region having $\Pr[Y = 1|x] = 1$ and points in the red region having $\Pr[Y = -1|x] = 1$. From the figure, it is clear that there exist two distinct vectors, w_1 and w_2, both of which define the lines (perpendicular to them) separating positive and negative examples.

The convex risk $\mathcal{R}$ is minimized by both of the sequences $(iw_1)_{i\geq 1}$ and $(iw_2)_{i\geq 1}$; moreover, the infimal risk is 0, which is not attained by any $w \in L_1(\mathcal{H}) = \mathbb{R}^2$, and every minimizing sequence has norms growing unboundedly.

Conceivably, minimizing logistic loss could lead one algorithm to follow the sequence $(iw_1)_{i\geq 1}$ and another to follow $(iw_2)_{i\geq 1}$. Both of these sequences converge in the $L_1(\mu)$ metric; their respective limit points, $\eta^{(1)}$ and $\eta^{(2)}$, are equal to 1, $1/2$, and 0 (for the positive class) on those points which have inner product, respectively, positive, 0, and negative to w_1 or w_2. Consequently, $\eta^{(1)} \neq \eta^{(2)}$. This shows that two different runs of logistic regression could give different probability estimates at some points. How then can Theorem 1.1 give a unique limit $\bar{\eta}$? The resolution is that Theorem 1.1 gives convergence in the $L_1(\mu)$ metric. In particular, w_1 and w_2 only disagree on the region between the two point clouds; this is a measure zero set, and thus $\eta^{(1)} = \eta^{(2)}$ μ-a.e.

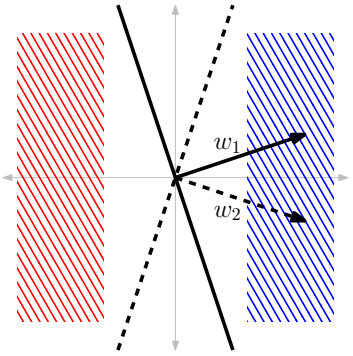

Figure 1: A well-separated classification problem.

Note that in this setting, it is also straightforward to prove an analog of the uniform deviation bounds of Theorem 1.2; indeed, applying either VC theory (Boucheron et al., 2005) or margin bounds (Schapire et al., 1997) will yield a bound that also lacks dependence on $\|w\|$. The distinction, however, is what both results say when applied to a sequence which does not achieve zero classification error. As will be shown in Proposition 1.3, the classification error of these sequences may be erratic, and therefore only loosely describes convergence behavior. On the other hand, Theorem 1.1 and Theorem 1.2 give a concrete object, $\bar{\eta}$, to which all minimizing sequences converge.

Classification errors and consistency. Let $\mathcal{R}_z(g) := \Pr[Y \neq \operatorname{sign}(g(X))]$ denote the classification error of any mapping $g : \mathcal{X} \to \mathbb{R}$, where $\operatorname{sign}(r) := \mathbf{1}[r \geq 0] - \mathbf{1}[r < 0]$. Recall that the signs of $\eta_w(\cdot, 1) - 1/2$ and Hw agree, which suggests that, because $\eta_w \to \bar{\eta}$ as provided by Theorems 1.1 and 1.2, there might be a relationship between $\left(\mathcal{R}_z(Hw_i)\right)_{i\geq 1}$ and $\left(\mathcal{R}_z(\bar{\eta}(\cdot, 1) - 1/2)\right)_{i\geq 1}$. However, convergence is stymied by the points where $\bar{\eta} = 1/2$, that is, the points where $\operatorname{sign}(\bar{\eta}(\cdot, 1) - 1/2)$ is discontinuous. The following result provides that, excluding this set, the desired convergence indeed occurs; in order to state it succinctly, further let $\eta_\mu(x, y) := \Pr[Y = y|x]$ denote the true conditional probability model, and $\mu_{\mathcal{X}}$ the marginal distribution along $\mathcal{X}$.

Proposition 1.3 *Suppose the setting of Theorem 1.1, and let $(w_i)_{i=1}^{\infty}$ be any sequence with $\eta_{w_i} \to \bar{\eta}$ in the $L_1(\mu)$ metric, and set $\Lambda := \{(x, y) : \bar{\eta}(x, y) = 1/2\}$. Then*

$$\limsup_{i\to\infty} \left| \mathcal{R}_z(Hw_i) - \mathcal{R}_z\left(\bar{\eta}(\cdot, 1) - \frac{1}{2}\right)\right| \leq \limsup_{i\to\infty} \underbrace{\left| \int_\Lambda \Big(2\eta_\mu(x, 1) - 1\Big) \mathbf{1}\left[\eta_{w_i}(x, 1) < \frac{1}{2}\right] d\mu_{\mathcal{X}}(x)\right|}_{\star}.$$

Moreover, there exist choices of $(\mu, \mathcal{H}, \ell)$ such that $\star > 0$ and the inequality is an equality.

The proposition implies that the difference between the classification error of $\bar{\eta}$ and that of η_{w_i} is bounded by $\mu(\bar{\eta} = 1/2)$ in the limit. The fact that the bound in the proposition can be tight, i.e., there is a gap between the classification risks even as $\eta_{w_i} \to \bar{\eta}$, implies that the classification risk cannot be easily used to show convergence of η_{w_i}. Similarly, as discussed with the example in Figure 1, any approach to the generalization analysis that bounds classification error, such as VC

theory, will be problematic since the classification error can behave erratically, as provided by the possibility of $\star > 0$ in Proposition 1.3.

Finally, recall the classical consistency results (Zhang, 2004; Bartlett et al., 2006), which may be summarized as follows. Let MF denote the set of all measurable functions. Then there exists a function $f_3 : \mathbb{R} \to \mathbb{R}_+$ with $f_3(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$ so that every $w \in L_1(\mathcal{H})$ satisfies

$$\mathcal{R}_z(w) - \inf_{f \in \mathrm{MF}} \mathcal{R}_z(f) \leq f_3\Big(\mathcal{R}(w) - \inf_{f \in \mathrm{MF}} \mathcal{R}(f)\Big),$$

where the last expression overloads $\mathcal{R}(f) = \int \ell(-yf(x))d\mu(x,y)$. As such, this result can be seen as a combination of Theorem 1.1 and Proposition 1.3 when $\mathrm{span}(\mathcal{H})$ is a rich family of functions (e.g., dense in MF). Consequently, the results of the present work can be seen as complementary, providing a specific convergence target $\bar{\eta}$ in the case of smaller $\mathrm{span}(\mathcal{H})$ (e.g., when $\inf_{w \in L_1(\mathcal{H})} \mathcal{R}(w) > \inf_{f \in \mathrm{MF}} \mathcal{R}(f)$), rather than a single-sided bound as above.

Outline. We close this introductory section with further notation. In Section 2, we construct $\bar{\eta}$ via convex duality, and sketch the proofs of Theorems 1.1 and 1.2 in Section 3. Many appendices collect further technical discussions and proof details.

Basic notation. Symbols defined in the preceding subsections—risk $\mathcal{R}$, excess risk $\mathcal{E}$, link function ϕ, conditional probability model η_w—will continue to be used in future sections. The weighting space $L_1(\mathcal{H})$ should be viewed as the L_1 space over the counting measure on elements of $\mathcal{H}$; since $h \in \mathcal{H}$ always has $\sup_x |h(x)| \leq 1$, it follows that $\sup_x |(Hw)(x)| \leq \|w\|_1$. Furthermore, in addition to the operator H, also let A denote the operator for which $(Aw)(x,y) = -y(Hw)(x)$, whereby

$$\mathcal{R}(w) = \int \ell\big((Aw)(x,y)\big)d\mu(x,y) = \int \ell(Aw)d\mu,$$

where the last form drops integration variables for succinctness.

We assume that μ can be *disintegrated* (Chang and Pollard, 1997) into a marginal measure $\mu_{\mathcal{X}}$ over $\mathcal{X}$, and a conditional probability $\eta_\mu(x,y) := \Pr[Y = y|x]$. Let $\mathcal{Z} := \mathcal{X} \times \{-1,+1\}$ be the set of all (x,y) pairs. Given any subset $C \subseteq \mathcal{Z}$, we define the intersection measure $\mu_C(S) := \mu(C \cap S)$ and conditional measure $\mu_{|C}$, where $\mu(C) > 0$ implies $\mu_C(S) = \mu_{|C}(S)\mu(C)$. We use a "hat" symbol to denote empirical measures, such as $\hat{\mu}$, $\hat{\mu}_C$, $\hat{\mu}_{|C}$. To avoid ambiguity, we sometimes write $\mathcal{R}(\cdot;\nu)$ and $\mathcal{E}(\cdot;\nu)$ to denote risk and excess risk when integration is over a measure ν.

Every loss $\ell : \mathbb{R} \to \mathbb{R}_+$ considered in this paper is a *classification loss*, meaning it is convex, non-decreasing, and satisfies $\ell(0) > 0$ and $\inf_{z \in \mathbb{R}} \ell(z) = 0$. The class of all such losses is denoted $\mathbb{L}$. The subset of these that are strictly convex and twice continuously differentiable (i.e., $\ell'' > 0$) is denoted $\mathbb{L}^{2+}$. The more restrictive class $\mathbb{L}^{2+}_{\mathrm{b}} \subseteq \mathbb{L}^{2+}$ will be defined in Section 3. For classification losses, which are not necessarily differentiable, we write $\ell'(z)$ to denote a fixed choice from the subgradient $\partial\ell(z)$; thus, a classification loss is described by a pair (ℓ, ℓ') satisfying $\ell'(z) \in \partial\ell(z)$.

2. Duality: The journey to the optimal conditional probability model $\bar{\eta}$

This section shows the existence of the optimal conditional probability model $\bar{\eta}$. The key challenge is the infinite dimensional setting, that is, the fact that the hypothesis space $\mathcal{H}$ and the sample space $\mathcal{Z}$ are infinite. To develop some intuition, we begin by studying the finite dimensional case.

2.1. Warm-up: Finite dimensional case

Assume for now that the hypothesis set is finite, $|\mathcal{H}| = d$, and the measure μ is uniform over n data points. Consider the problem of optimizing exponential loss over this measure, i.e.,

$$\inf_{w\in\mathbb{R}^d}\left[\sum_{i=1}^{n} e^{-y_i(Hw)(x_i)}\right] \quad . \tag{2}$$

The conditional model for the exponential loss is

$$\eta_w(x,y) = \frac{e^{y(Hw)(x)}}{e^{y(Hw)(x)} + e^{-y(Hw)(x)}} \quad . \tag{3}$$

Recalling the example from Figure 1, note how easily the infimum to Eq. (2) may fail to be attained. In particular, if there exists $\hat{w} \in \mathbb{R}^d$ defining a hyperplane which strictly separates the positive and negative examples, then the sequence $(j\hat{w})_{j\geq 1}$ achieves zero risk in the limit, whereas every element $w \in \mathbb{R}^d$ achieves a positive risk. On the other hand, $\eta_{j\hat{w}}(x_i,y_i) \to 1$ as $j\to\infty$. So in this case, $\bar{\eta}$, which needs to be defined only over the examples (x_i,y_i), is described by $\bar{\eta}(x_i,y_i) = 1$.

Similar to other studies of risk minimization stymied by the problem of missing minimizers (Collins et al., 2002), we consider the convex dual to Eq. (2). The dual of loss minimization of a linear model is the problem of maximizing entropy subject to constraints, where different losses yield different kinds of entropy (Collins et al., 2002; Altun and Smola, 2006). The dual of Eq. (2) is

$$\max_{q\in\mathbb{R}^n_+}\left[\sum_{i=1}^{n}\left(-q_i\ln q_i + q_i\right)\right] \quad \text{s.t.} \quad \sum_{i=1}^{n} q_i\big(y_i h(x_i)\big) = 0 \text{ for all } h\in\mathcal{H} \quad . \tag{4}$$

The objective on the left is an unnormalized entropy of the dual variable vector q, representing an unnormalized reweighting of examples. The unnormalized entropy is being maximized over the set of reweightings, which satisfy "decorrelation" constraints on the right. Specifically, the constraints require that the reweighting q be uncorrelated with every hypothesis, making the reweighted prediction problem as hard as possible. Note that $q = 0$ is always feasible, but the unnormalized entropy pushes the solution away from zero whenever feasible (the slope of entropy at zero is $-\infty$ (Lemma E.1.v)). Theorem 2.1 shows that the dual maximum is always attained, unlike the primal minimum. However, if both the primal maximum $\bar{w}$ and dual maximum $\bar{q}$ are attained, then $\bar{q}_i = \exp\big(-y_i(H\bar{w})(x_i)\big)$. For a general differentiable loss ℓ, the optimality conditions yield $\bar{q}_i = \ell'\big(-y_i(H\bar{w})x_i\big)$. If there is any example j such that $x_j = x_i$, but the label is flipped ($y_j = -y_i$), then we can rewrite $\bar{q}_j$ as $\bar{q}_j = \exp\big(y_i(H\bar{w})(x_i)\big)$ for exponential loss, and $\bar{q}_j = \ell'\big(y_i(H\bar{w})(x_i)\big)$ for a general loss. Let $-i$ denote such an index j if it exists. Contrasting the definition of $\eta_{\bar{w}}$ in Eq. (3) with the optimality condition for $\bar{q}$ suggests defining

$$\bar{\eta}(x_i,y_i) = \begin{cases} \bar{q}_{-i}/\left(\bar{q}_{-i} + \bar{q}_i\right) & \text{if } \bar{q}_i > 0, \\ 1 & \text{if } \bar{q}_i = 0, \end{cases}$$

where in the absence of the example with the flipped label, define $\bar{q}_{-i} = \ell'\big(-(\ell')^{-1}(\bar{q}_i)\big)$ to emulate such an example; for exponential loss, $\bar{q}_{-i} = 1/\bar{q}_i$. The value of $\bar{\eta}$ for $\bar{q}_i = 0$ is obtained by taking the limit $\bar{q}_i \to 0$ (i.e., $\bar{q}_{-i}\to\infty$ for exponential loss). The next section shows that this $\bar{\eta}$ is the correct limit object, even for an infinite sample space and an infinite hypothesis set.

2.2. Infinite dimensional case

Before constructing $\bar{\eta}$ and proving Theorem 1.1, we establish an infinite dimensional duality result similar to the finite dimensional result from Section 2.1. In the primal, we now minimize an integral rather than a sum. In the dual, we optimize over unnormalized densities over $\mathcal{Z}$. Recall that the linear map A returns functions on $\mathcal{Z}$ such that $(Aw)(x,y) = -y(Hw)(x)$. Formally, we seek the following duality result:

$$\inf_{w \in L_1(\mathcal{H})} \left[\int \ell(Aw) d\mu \right] = \max_{q \in \mathcal{Q}} \left[- \int \ell^* \Big(q(x,y) \Big) d\mu(x,y) \right] \quad \text{s.t.} \int q(x,y) \Big(y h(x) \Big) d\mu(x,y) = 0 \text{ for all } h \in \mathcal{H} \tag{5}$$

where $\ell^*(s) := \sup_r [rs - \ell(r)]$ is the conjugate of ℓ (see Appendix A). For example, when ℓ denotes the exponential loss, we find that $\ell^*(s) = s \ln s - s$ for $s \geq 0$ and $\ell^*(s) = \infty$ for $s < 0$, giving rise to the non-negativity constraint on q and the dual objective we already saw in Eq. (4).

A crucial technical question is the choice of $\mathcal{Q}$, i.e., the set that q is selected from. Following the intuition of Section 2.1, the goal is to construct $\bar{\eta}(x,y) = \bar{q}(x,-y) / \big(\bar{q}(x,-y) + \bar{q}(x,y) \big)$. The space $\mathcal{Q}$ should be large enough to allow construction of any conditional probability distribution η for μ_X. To achieve this, it suffices to make sure that all measures which are absolutely continuous with respect to μ have their densities included in $\mathcal{Q}$. In fact, our set can be slightly smaller: it just needs to include all densities for which the dual objective, i.e., the integral $\int \ell^*(q) d\mu$, is finite.

One candidate class of functional spaces is $L_p(\mu)$, where $p \geq 1$. These are Banach spaces of measurable functions with the norm defined by $\|f\|_p = \left(\int |f|^p \, d\mu \right)^{1/p}$. The space $L_p(\mu)$ contains all measurable functions with $\|f\|_p < \infty$. However, in our setting, we instead want to place restrictions on the allowed functions q based on the integral $\int \ell^*(q) d\mu$ rather than $\int |q|^p d\mu$. Therefore, instead of working with $L_p(\mu)$ spaces, we work with their generalization called *large Orlicz spaces* (Léonard, 2007, and Appendix B), which allows us to tailor the set $\mathcal{Q}$ to ℓ^*.

In detail, the construction of a large Orlicz space begins with a non-negative convex function $\theta : \mathbb{R} \to [0, \infty]$ symmetric around zero (i.e., $\theta(r) = \theta(|r|)$), not identical to zero (i.e., $\theta(r) \to \infty$ as $r \to \infty$, by convexity), and with $\theta(0) = 0$. This function θ serves the same role as the p-th power function in the construction of $L_p(\mu)$. The conditions that we place on θ make it possible to define "the unit ball" of functions, analogous to the unit ball in $L_p(\mu)$, namely

$$\mathcal{B} := \left\{ f \text{ measurable} : \int \theta\big(f(z)\big) d\mu(z) \leq 1 \right\} .$$

This set is then used to define the norm $\|f\|_\theta = \inf\{r \geq 0 : f \in r\mathcal{B}\}$, where the norm equals ∞ if f is outside the scaled ball $r\mathcal{B}$ for all $r \geq 0$. The *large Orlicz space* $L_\theta(\mu)$ is defined to contain all measurable functions with $\|f\|_\theta < \infty$. For $p \geq 1$, the choice $\theta(s) = |s|^p$ recovers the $L_p(\mu)$ spaces. (See Appendix B for further background.)

Now we are ready to answer what the space $\mathcal{Q}$ should be. Following the construction of (Léonard, 2008), we begin by introducing a symmetrized version of the loss ℓ with the first-order Taylor expansion at zero removed:

$$\beta(s) := \max \left\{ \ell(s) - \Big(\ell(0) + s\ell'(0) \Big), \; \ell(-s) - \Big(\ell(0) + (-s)\ell'(0) \Big) \right\} . \tag{6}$$

It turns out that the Orlicz space $L_{\beta^*}(\mu)$, derived from the conjugate β^*, satisfies our desideratum on $\mathcal{Q}$: it contains all the densities with respect to μ whose dual objective is finite (see Lemma G.1.iii).

The next theorem spells out the duality result of Eq. (5) with a more succinct representation of constraints via *adjoint* $A^\top$ of the operator A. The adjoint is a generalization of the matrix transpose. The adjoint $A^\top$ is a linear operator which maps q into a linear function on $L_1(\mathcal{H})$ defined by $(A^\top q)(w) = \int (Aw)(z)\, q(z)\, d\mu$. The constraint of Eq. (5) is equivalent to requiring $(A^\top q)(w) = 0$ for all w, i.e., $A^\top q$ is required to be the zero of the vector space of linear functions on $L_1(\mathcal{H})$. Thus, the constraint can be written as $A^\top q = 0$, highlighting the fact that it is a linear constraint on q.

Apart from the duality result, the theorem also enumerates several important properties of the dual optimum, which are relevant for the construction of $\bar{\eta}$ in Definition 2.2 below. Properties (i) and (ii) show that $\bar{\eta}$ is a well-defined conditional probability. Property (iii) implies that $\bar{\eta}(x,y) = \eta_{\bar{w}}(x,y) = \phi(y(H\bar{w})(x))$ when the primal optimum exists and the loss is differentiable. Property (iv) looks more technical: it implies that when the primal optimum $\bar{w}$ does not exist, $\bar{h}(x) := (\ell')^{-1}(\bar{q}(x,-1))$ can serve a similar role as $H\bar{w}$, because $\bar{\eta}(x,y) = \phi(y\bar{h}(x))$; indeed, we use this construction of $\bar{h}$ in Section 3.1.

Theorem 2.1 *Let finite measure μ over $\mathcal{Z}$, hypotheses $\mathcal{H}$, and loss function $\ell \in \mathbb{L}$ be given, with β defined by Eq. (6). Then*

$$\inf_{w \in L_1(\mathcal{H})} \left[\int \ell(Aw) d\mu \right] \quad = \quad \max_{q \in L_{\beta^*}(\mu):\ A^\top q = 0} \left[- \int \ell^*(q) d\mu \right] \ . \tag{7}$$

A dual optimum $\bar{q}$ always exists, and can be chosen to satisfy the following, μ-a.e. over (x,y):

(i) $\bar{q}(x,y) \geq 0$.
(ii) $\bar{q}(x,y) + \bar{q}(x,-y) > 0$.
(iii) $\bar{q}(x,y) \in \partial \ell(A\bar{w})(x,y)$ *where $\bar{w}$ is a primal optimum (if it exists).*

Furthermore,

(iv) *If $\ell \in \mathbb{L}^{2+}$, then $(\ell')^{-1}\big(\bar{q}(x,y)\big) = -(\ell')^{-1}\big(\bar{q}(x,-y)\big)$, μ-a.e. over all (x,y) for which $(\ell')^{-1}$ is defined at both $\bar{q}(x,y)$ and $\bar{q}(x,-y)$.*
(v) *If ℓ is differentiable, then $\bar{q}$ is unique (up to μ-null sets).*

Using part (v), we obtain that the following defines a unique $\bar{\eta}$ (up to μ-null sets):

Definition 2.2 *Let $\ell \in \mathbb{L}$ be differentiable and $\bar{q}$ be the dual optimum satisfying conditions (i) and (ii) of Theorem 2.1. We define the optimal conditional model $\bar{\eta}$ as*

$$\bar{\eta}(x,y) = \frac{\bar{q}(x,-y)}{\bar{q}(x,-y) + \bar{q}(x,y)} \ . \tag{8}$$

This is the $\bar{\eta}$ that appears in Theorem 1.1. This theorem will be proved in the next section.

3. Convergence and generalization via easy and difficult sets

We saw in Section 2.1 that the conjugate ℓ^* of the exponential loss has an infinite slope at zero; the same turns out to be true for all losses in $\mathbb{L}^{2+}$ (Lemma E.1.v). Informally, this means that the dual optimization avoids setting $\bar{q} = 0$ unless forced to do so by the decorrelation constraint $A^\top \bar{q} = 0$. We will see that this distinction between the set of points where $\bar{q} = 0$ and the set where $\bar{q} > 0$ is fundamentally important to the analysis, a fact seen before in the analysis of boosting (Mukherjee et al., 2011; Telgarsky, 2012, 2013). We call these two sets of points "easy" and "difficult" (respectively) for reasons which we illustrate on an example.

An example. Consider the example in Figure 2, which builds on the example from Figure 1. In addition to the two well-separated regions of positive and negative examples, we now add an alternating sequence of positive and negative point masses along the line γ orthogonal to the weight vector w_1. Each weight vector $w \in \mathbb{R}^2$ represents a linear predictor returning the inner product $x \mapsto w \cdot x$. The *margin* of a data point (x, y) with respect to this predictor is $y(w \cdot x)$. The decorrelation constraint (see Eq. (5)) requires that the weighted margin of every hypothesis (and of every linear combination) according to the density q is equal to zero. The predictor described by w_1 gives a positive margin to all points in the two separated regions (the easy set) and zero margin to those along the line γ (the difficult set). Hence, any q satisfying the decorrelation constraint must equal zero over these two regions. On the other hand, because the point masses along γ are antisymmetric around zero, each of them can receive the density $q(x, y) = \bar{s}$ where $\bar{s}$ is a minimizer of ℓ^* (it always exists by Lemma E.1.i).

In the primal, the sequence $(iw_1)_{i\geq 1}$ still minimizes the risk as follows. First, the risk in the two regions goes to zero. Next, the risk of any weight vector w over points along γ is only a function of the projection of w onto γ. Since the masses along γ are antisymmetric and the loss function is convex (and increasing as the prediction is more wrong), the projection needs to be at the origin to minimize the risk along γ. This is exactly the case for iw_1 by orthogonality.

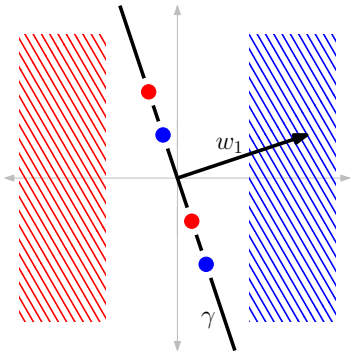

Figure 2: Easy and difficult sets.

If the example were to be slightly perturbed, so that the point masses would still lie on γ in an alternating pattern (but not antisymmetric), a minimizing sequence would take the form $(\hat{w} + iw_1)_{i\geq 1}$ where $\hat{w} \in \gamma$ would be the minimizer of the risk of the points along γ. Because of the alternating pattern such a minimizer would be bound to exist.

Preliminaries. Several aspects of the example carry over to the general setting. First, it can be shown that the risk on points where $\bar{q} = 0$ converges to zero when the primal is minimized, that is, a perfect classification is achieved. Therefore, we call this set of points "easy". Second, the points where $\bar{q} > 0$ cannot be further "separated" in the sense that any w under which some non-null measure of these points receives a positive margin also yields a non-null measure of points with a negative margin. We call this set "difficult".

Definition 3.1 *Given a finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, and a dual optimum $\bar{q}$ satisfying the conditions of Theorem 2.1, the* difficult set *is defined as $\mathcal{D} := \{z \in \mathcal{Z} : \bar{q}(z) > 0\}$. Its complement $\mathcal{D}^c$ is called the* easy set.

The next lemma (and the following corollary) show that, similar to the example, all of the risk is in fact due to the difficult set. The lemma proves equality of the dual objectives for μ and the restricted measure $\mu_{\mathcal{D}}$, and furthermore that $\bar{q}$ is feasible and optimal for both problems. The corollary highlights the implications in the primal, that by optimizing the risk on μ, we optimize the risk on the difficult set, and drive the risk on the easy set to zero. For technical reasons, both results are stated for *supersets* of difficult sets.

Lemma 3.2 *Given a finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, a difficult set $\mathcal{D}$ and an associated dual optimum $\bar{q}$, let D be an arbitrary (measurable) superset of the difficult set: $\mathcal{D} \subseteq D$. Then the*

dual optimal values for μ and μ_D are equal:

$$\max_{q\in L_{\beta^*}(\mu):\, A^\top q=0} \left[-\int \ell^*(q)d\mu\right] \;=\; \max_{q\in L_{\beta^*}(\mu_D):\, A^\top q=0} \left[-\int_D \ell^*(q)d\mu\right] \quad .$$

The general dual optimum $\bar{q}$ is feasible for both problems and attains both maxima. Moreover, if $\bar{q}_D$ is a dual optimum for μ_D, then $\hat{q}(z) := \bar{q}_D(z)\mathbf{1}[z \in D]$ is also a dual optimum for both problems.

Corollary 3.3 *Let D be a superset of a difficult set, $\mathcal{D} \subseteq D$. Then $\mathcal{E}(w;\mu_D) \leq \mathcal{E}(w)$ and $\mathcal{R}(w;\mu_{D^c}) \leq \mathcal{E}(w)$ for all $w \in L_1(\mathcal{H})$.*

We wrap up this section by defining the class $\mathbb{L}_{\mathrm{b}}^{2+}$ appearing in our main results. While the class may appear restrictive, it contains the logistic and exponential losses by Proposition E.6:

Definition 3.4 *The class $\mathbb{L}_{\mathrm{b}}^{2+} \subseteq \mathbb{L}^{2+}$ consists of strictly convex, twice continuously differentiable classification losses ℓ, which in addition satisfy the following conditions:*

(i) *The link function ϕ, derived from ℓ as in Eq. (1), is Lipschitz-continuous with constant L_ϕ.*
(ii) *For some $c_\ell > 0$, the derivative ℓ' satisfies $\ell'(r) \leq c_\ell \ell(r)$ whenever $r \leq 0$.*
(iii) *For every finite measure μ over $\mathcal{Z}$, there exists $c_{\ell,\mu} \geq 0$ with $\|f\|_\beta \leq c_{\ell,\mu} \int \ell(f)d\mu$ for every measurable $f : \mathcal{Z} \to \mathbb{R}_+$.*

3.1. Proof outline for Theorem 1.1

Recall that our goal is to show that risk minimization yields convergence of η_w to $\bar{\eta}$. First consider the easy set $\mathcal{D}^c$. By Corollary 3.3, minimizing $\mathcal{R}(w)$, i.e., taking $\mathcal{E}(w)$ to zero, leads to $\mathcal{R}(w;\mu_{\mathcal{D}^c})$ becoming arbitrarily small. This in turn means that most predictions $(Hw)(x)$ will not only have the correct sign, but will also have a large margin. This observation can be used to obtain the following bounds on a partition of the easy set $\mathcal{D}^c$ into two sets: S_r and $\mathcal{D}^c \setminus S_r$. The bound on $\mu(S_r)$ is also a bound on $\int_{S_r} |\bar{\eta} - \eta_w| d\mu$ because $|\bar{\eta} - \eta_w| \leq 1$. Thus, together these bound $\int_{\mathcal{D}^c} |\bar{\eta} - \eta_w| d\mu$.

Lemma 3.5 *Given a finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, and a difficult set $\mathcal{D}$, let D be an arbitrary (measurable) superset of the difficult set: $\mathcal{D} \subseteq D$. Let any $w \in L_1(\mathcal{H})$ and $r > 0$ be given, and define $S_r := \{z \in D^c : \ell\big((Aw)(z)\big) \geq r\}$. Then:*

(i) $\mu(S_r) \leq \mathcal{R}(w;\mu_{D^c})/r \leq \mathcal{E}(w)/r$,
(ii) $\int_{D^c\setminus S_r} |\bar{\eta} - \eta_w|\, d\mu \leq r\mu(D^c \setminus S_r) \max\{1/\ell(0),\ c_\ell/\ell'(0)\}$ *if* $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$.

It remains to control η_w over $\mathcal{D}$. As mentioned earlier, the decorrelation constraint implies that the difficult set $\mathcal{D}$ cannot be "separated" in the sense that any w under which some subset of $\mathcal{D}$ with a positive measure μ has a positive margin (i.e., correct predictions), also yields a positive measure of points in $\mathcal{D}$ with a negative margin (i.e., incorrect predictions). Since the loss is increasing over negative margins, this structure implies that the risk over $\mathcal{D}$ has a minimizer over each one-dimensional subspace (similar reasoning to the example of Figure 2). This one-dimensional property can be used in finite dimensions to argue that the risk must have a minimizer over the difficult set, and we pursue this line of reasoning in Section 3.2. But here, we need an alternative approach.

As discussed in Theorem 2.1.iv, if $\ell \in \mathbb{L}^{2+}$, then $\bar{\eta}(x,y) = \phi(-\bar{f}(x,y))$ with $\bar{f}(x,y) = (\ell')^{-1}(\bar{q}(x,y))$ whenever $(\ell')^{-1}$ is defined for both $\bar{q}(x,y)$ and $\bar{q}(x,-y)$. Fortunately, this can be shown to hold μ-a.e. over $\mathcal{D}$. Thus, over $\mathcal{D}$, we can write $|\bar{\eta} - \eta_w| = \big|\phi(-\bar{f}) - \phi(-Aw)\big|$. The next lemma uses a second-order Taylor expansion at $\bar{f}$ to further derive a bound on this difference.

In Lemma 3.6, we split the difficult set into four subsets and we either bound their mass, which in turn bounds the integral of $|\bar{\eta} - \eta_w|$, or directly bound the integral. The integral is controlled directly over the subset U by the mentioned Taylor bound, and so it requires the bounds on the range of Aw and $\bar{f}$ (via $\bar{q}$), and a corresponding lower bound τ on the second derivative. The subset S_+ contains points with a large loss, so its mass is controlled by the risk. The control of the subset S_- is the most technical. The set includes points where the predictions are correct, but the density $\bar{q}$ is large. The bound is based on the decorrelation constraint as well as property (iii) in Definition 3.4. All three bounds depend on w only via its risk; this is indeed key to establishing Theorem 1.1. The set V needs to be controlled separately.

Lemma 3.6 *Given a finite measure μ with $\mu(\mathcal{Z}) \leq 1$, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$, a difficult set $\mathcal{D}$ and an associated dual optimum $\bar{q}$, let a weighting $w \in L_1(\mathcal{H})$ be given, along with scalars $c_1 > 0$, $c_2 > 0$, $c_3 > c_2$, and $\tau := \min\{\inf_{|z| \leq c_1} \ell''(z), \inf_{z \in [c_2, c_3]} \ell''((\ell^*)'(z))\}$. Define the following sets:*

$$U := \{z \in \mathcal{D} : |(Aw)(z)| \leq c_1 \text{ and } c_2 \leq \bar{q}(z) \leq c_3\}, \quad S_+ := \{z \in \mathcal{D} : (Aw)(z) > c_1\},$$
$$S_- := \{z \in \mathcal{D} : (Aw)(z) < -c_1 \text{ and } \bar{q}(z) \geq c_2\}, \qquad V := \{z \in \mathcal{D} : \bar{q}(z) < c_2 \text{ or } \bar{q}(z) > c_3\}.$$

Then $\mathcal{D} = U \cup S_+ \cup S_- \cup V$,

$$\mu(S_+) \leq \frac{\mathcal{R}(w)}{c_1 \ell'(0)}, \quad \mu(S_-) \leq \frac{2c_{\ell,\mu} \|\bar{q}\|_{\beta^*} \mathcal{R}(w)}{c_1 c_2}, \quad \int_U |\bar{\eta} - \eta_w| \, d\mu \leq L_\phi \sqrt{\frac{2\mathcal{E}(w; \mu_{\mathcal{D}})}{\tau}}.$$

To prove Theorem 1.1 from here, first split $\int |\eta_w - \bar{\eta}| d\mu$ along $\mathcal{D}$ and $\mathcal{D}^c$, and apply Lemma 3.5 and Lemma 3.6 to the two pieces; the goal is to show that all terms go to zero as $\mathcal{E}(w) \to 0$. In the terms resulting from Lemma 3.5, this is handled by the choice $r := \sqrt{\mathcal{E}(w)}$. Similarly, it is possible (although considerably more challenging) to balance c_1, c_2, c_3, τ arising from Lemma 3.6.

3.2. Proof outline for Theorem 1.2

In this section we sketch the proof of the generalization bound from the introduction (Theorem 1.2). Unlike the foregoing results, here we assume that the hypothesis space is finite, $|\mathcal{H}| = d$.

Similar to Section 3.1, the proof treats the easy set and the difficult set separately. On the easy set, where zero risk is possible in the limit, linear predictors actually achieve zero *classification error* when viewed as half-space classifiers. Finite dimension d then implies a finite VC dimension and the corresponding generalization bound. In the remainder, we only focus on the difficult set.

We build on the fact that on the difficult set $\mathcal{D}$ the risk is eventually increasing along any direction which lies in the "span" of $\mathcal{D}$ (similar to the example of Figure 2). In the finite dimension d, this will imply a bound on the norm of the optimizer of risk over $\mathcal{D}$, and also enable the application of Rademacher complexity to obtain a generalization bound.

We begin with a specific lower bound in each direction w within the "span". The bound is obtained by integrating over all points with a negative margin, i.e., $(Aw)(z) > 0$. Because of the lack of separators over $\mathcal{D}$, the bound is non-zero. Taking an infimum over all directions yields a uniform bound called *balance*. While the following definition is written for any measure μ, it is going to be primarily applied with $\mu_{\mathcal{D}}$ substituted for μ:

Definition 3.7 *The* balance *associated with hypotheses $\mathcal{H}$, $|\mathcal{H}| = d$, and measure μ is defined as* $\mathrm{Bal}(\mu) := \inf\left\{\int |(Aw)(z)|_+ \, d\mu(z) : w \in \mathrm{Ker}(\mu)^\perp, \|w\|_1 = 1\right\}$, *where $|s|_+ := \max\{s, 0\}$ denotes the non-negative part, and* $\mathrm{Ker}(\mu) := \left\{w \in \mathbb{R}^d : (Aw)(z) = 0, \ \mu\text{-a.e. over } z\right\}$ *denotes the subspace of $\mathbb{R}^d$ with no effect on risk under μ.*

The "span" corresponds to the orthogonal complement of the kernel $\mathrm{Ker}(\mu)$. In the example of Figure 2, the difficult set consisted of the points on the line γ, and the kernel $\mathrm{Ker}(\mu_{\mathcal{D}})$ was the subspace spanned by the vector w_1, which had no effect on the risk over points on γ. The only interesting directions from the perspective of this risk were in the orthogonal complement $\mathrm{Ker}(\mu_{\mathcal{D}})^{\perp}$.

In finite dimension d, we obtain that $\mathrm{Bal}(\mu_{\mathcal{D}}) > 0$ whenever $\mu(\mathcal{D}) > 0$ (Proposition J.4). This yields a non-trivial risk bound from the definition of balance, using the fact that $\ell(r) \geq \ell(0) + r\ell'(0) \geq r\ell'(0)$ (by convexity and non-negativity of ℓ):

$$\mathcal{R}(w;\mu_{\mathcal{D}}) \geq \int_{Aw>0} \ell(Aw) d\mu_{\mathcal{D}} \geq \int_{Aw>0} \ell'(0)(Aw) d\mu_{\mathcal{D}} \geq \ell'(0)\|w\|_1 \mathrm{Bal}(\mu_{\mathcal{D}}).$$

Rearranging, we also obtain a norm bound $\|w\|_1 \leq \mathcal{R}(w)/(\ell'(0)\mathrm{Bal}(\mu_{\mathcal{D}}))$, which enables the use of Rademacher complexity in the analysis of generalization on $\mathcal{D}$.

A less obvious consequence is that for a given finite hypothesis class $\mathcal{H}$ and measure μ, there exists a maximal difficult set. This difficult set, common to the entire class $\mathbb{L}^{2+}$, is called the *canonical difficult set* $\mathcal{D}_\star$ (for concreteness, we define it for $\ell = \exp$). Informally, its existence follows from the property shared by all losses $\ell \in \mathbb{L}^{2+}$ that $(\ell^*)'(s) \uparrow \infty$ as $s \downarrow 0$ (Lemma E.1.v); consequently, the optimization prevents $\bar{q}$ from taking on the value zero unless forced by constraints, and thus yields the largest possible difficult set:

Definition 3.8 *For a finite measure μ and a hypothesis set with $|\mathcal{H}| < \infty$, the* canonical difficult set $\mathcal{D}_\star$ *is defined as any difficult set associated with $\ell = \exp$.*

Proposition 3.9 *Given a finite measure μ, a hypothesis set with $|\mathcal{H}| < \infty$, and the corresponding canonical difficult set $\mathcal{D}_\star$, we have:*

(i) *For any $\ell \in \mathbb{L}$ and any corresponding difficult set $\mathcal{D}$, we have $\mathcal{D} \subseteq \mathcal{D}_\star$ μ-a.e.*
(ii) *For any $\ell \in \mathbb{L}^{2+}$ and any corresponding difficult set $\mathcal{D}$, we have $\mathcal{D} = \mathcal{D}_\star$ μ-a.e.*

We finish this section with the Rademacher complexity style bound on the excess risk over the canonical difficult set $\mathcal{D}^*$, based on the norm bound implied by the balance. The key insight is that the quantities in the bound depend on w only through the empirical risk $\mathcal{R}(w; d\widehat{\mu}_{|\mathcal{D}_\star})$. Theorem 1.2 is then proved by splitting $\int |\eta_w - \bar{\eta}| d\mu$ along $\mathcal{D}_\star$ and $\mathcal{D}_\star^c$, and controlling the pieces by a combination of Lemma 3.5 with the VC style bound (Lemma J.9) used to select r, and Lemma 3.6 with the scalars chosen via Lemma 3.10.

Lemma 3.10 *Let probability measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| = d$, loss function $\ell \in \mathbb{L}$, subgradient $\bar{s} \in \partial\ell(0)$, and a canonical difficult set $\mathcal{D}_\star$ with $\mu(\mathcal{D}_\star) > 0$ be given. Set $\tau(r) := \inf_{|z|\leq r} \ell''(z)$, $\mathrm{Bal}_\star := \mathrm{Bal}(\mu_{|\mathcal{D}_\star})$ and let $B_w := 2 + \left[\ell(0) + 2\mathcal{R}(w; d\widehat{\mu}_{|\mathcal{D}_\star})\right]/(\bar{s}\mathrm{Bal}_\star)$, and $n \geq 256\ln(8d/\delta)/\mathrm{Bal}_\star^2$. Then with probability at least $1 - 4\delta$ over a draw from $\mu_{|\mathcal{D}_\star}$ of size n, the following statements hold simultaneously for every $w \in L_1(\mathcal{H})$:*

(i) $|(Aw)(z)| \leq B_w$ *for μ-a.e. and $\widehat{\mu}$-a.e. $z \in \mathcal{D}_\star$.*
(ii) $\mathcal{E}(w, \mu_{|\mathcal{D}_\star}) \leq \mathcal{E}(w, \widehat{\mu}_{|\mathcal{D}_\star}) + 10\ell(2B_w)\sqrt{\ln(8dB_w^2/\delta)/n}.$
(iii) $\mathcal{E}(w, \mu_{|\mathcal{D}_\star}) \leq 2\mathcal{E}(w, \widehat{\mu}_{|\mathcal{D}_\star}) + \dfrac{1024\ell'(2B_w)^2 \ln(8dB_w^2/\delta)}{n\mathrm{Bal}_\star^2 \tau(B_w)}$ *if $\ell \in \mathbb{L}_{\mathfrak{b}}^{2+}$.*

Acknowledgments

The authors thank Rastislav Telgársky for pointing out that the topology should always be adapted to the problem at hand; may he rest in peace.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. 2006.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, Nov 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 08 2005. doi: 10.1214/009053605000000282.

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

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Joseph T. Chang and David Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3): 287–317, 1997.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2):337–407, 2000.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *FOCS*, 2006.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

Michael Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT Press, 1994.

Christian Léonard. Orlicz spaces. http://www.cmap.polytechnique.fr/~leonard/papers/orlicz.pdf, 2007. Accessed 2015-04-28.

Christian Léonard. Minimization of entropy functionals. *J. Math. Anal. Appl.*, 346:183–204, 2008.

Kfir Levy, Elad Hazan, and Tomer Koren. Logistic regression: Tight bounds for stochastic and online optimization. In *COLT*, 2014.

Indraneel Mukherjee, Cynthia Rudin, and Robert Schapire. The convergence rate of AdaBoost. In *COLT*, 2011.

R. Tyrrell Rockafellar. Integrals which are convex functionals I. *Pacific J. Math.*, 24:525–539, 1968.

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

R. Tyrrell Rockafellar. *Conjugate Duality and Optimization*. SIAM Publications, 1974.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Nathan Srebro, and Karthik Sridharan. Fast rates for regularized objectives. In *NIPS*, 2008.

Matus Telgarsky. A primal-dual convergence analysis of boosting. *JMLR*, 13:561–606, 2012.

Matus Telgarsky. Boosting with the logistic loss is consistent. In *COLT*, 2013.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

Appendix A. Convex analysis in Banach spaces

This appendix covers convex analysis results for functional spaces. It is based on Rockafellar (1974) and Rockafellar (1968).

Banach spaces. A *Banach space* is a complete normed vector space. The space $\mathbb{R}^n$ with the Euclidean norm is a Banach space. Given a measure μ on $\mathcal{Z}$ and $p \geq 1$, the Banach space $L_p(\mu)$ consists of all measurable functions $f : \mathcal{Z} \to \mathbb{R}$ with the finite norm $\|f\|_p := \left(\int |f|^p \, d\mu\right)^{1/p}$.

The analog of an inner product for Banach spaces is a *pairing*. Given two Banach spaces U and V, their pairing is described by a bilinear form $U \times V \to \mathbb{R}$, denoted $\langle u, v\rangle$. Thus, each $u \in U$ describes a linear map $v \mapsto \langle u, v\rangle$ on V and vice versa. Each Banach space is endowed with the topology implied by its norm, but other topologies are possible. We say that the topologies on U and V are *compatible* with the pairing if the linear functions described by $u \in U$ and $v \in V$ are continuous, and if they comprise all continuous linear functions on V and U, respectively. A Euclidean space $\mathbb{R}^n$ with the norm topology is compatibly paired with itself via standard inner product. Given $1 < p, q < \infty$ such that $1/p + 1/q = 1$, the spaces $L_p(\mu)$ and $L_q(\mu)$ with their norm topologies are a compatible pairing with the bilinear form $\langle f, g\rangle = \int fg \, d\mu$. One construction of compatible pairings begins with a Banach space U under norm topology, then takes its *topological dual* U', i.e., the space of all continuous linear functions on U, and endows U' with the *weak* topology*. In the rest of the paper, when we talk about "paired Banach spaces" we assume that they have been endowed with compatible topologies.

Convexity, conjugacy, subgradients. Given a Banach space U, a function $F : U \to (-\infty, \infty]$ is called *proper* if it is not equal to ∞ everywhere. The set of points where F is finite is called its *domain* and denoted $\operatorname{dom} F$. The *epigraph* of F is the set of points above the graph of the function $\{(u,t) : u \in U, t \in \mathbb{R}, t \geq F(u)\}$. The function F is called *convex* if its epigraph is convex. It is called *closed* if its epigraph is closed.

Let U and V be paired Banach spaces. Let $F : U \to (-\infty, \infty]$ be a closed proper convex function. The *conjugate* of F is defined by $F^*(v) := \sup_{u \in U} \big[\langle u, v\rangle - F(u)\big]$. It is also a closed proper convex function and $F^{**} = F$ (Theorem 5 of Rockafellar, 1974). From the definition of a conjugate, we get *Fenchel's inequality*

$$F(u) + F^*(v) \geq \langle u, v\rangle \ .$$

The *subgradient* of F at u is the set $\partial F(u) := \{v \in V : F(u') \geq F(u) + \langle u' - u, v\rangle \text{ for all } u' \in U\}$. For a closed proper convex function F, the following statements are equivalent (Corollary 12A and the foregoing discussion of Rockafellar, 1974) (first-order optimality for conjugates):

(i) $F(u) + F^*(v) = \langle u, v\rangle$,
(ii) $v \in \partial F(u)$,
(iii) $u \in \partial F^*(v)$.

Integrals as convex functionals. Consider a finite measure μ on $\mathcal{Z}$, and assume we are given a pairing of Banach spaces U and V via bilinear form $\langle u, v\rangle = \int u(z)v(z)\, d\mu(z)$, i.e., U and V are subsets of measurable functions on $\mathcal{Z}$. Let $f : \mathbb{R} \to (-\infty, \infty]$ be a closed proper convex function. We study properties of the function F on U defined by the integral

$$F(u) = \int f\big(u(z)\big)\, d\mu(z) \ .$$

To establish its closedness and study conjugacy we need the following definition, adapted from Rockafellar (1968) for the case of a finite measure μ:

Definition A.1 *We say that a Banach space of measurable functions on $\mathcal{Z}$ is decomposable with respect to a finite measure μ if the following conditions hold:*

(i) *U contains every bounded measurable function from $\mathcal{Z}$ to $\mathbb{R}$.*
(ii) *If $u \in U$ and E is a measurable set, then U contains $u \cdot \mathbf{1}_E$ where $\mathbf{1}_E$ is the indicator of the set E.*

The following proposition is a rephrasing of the corollary on page 534 of Rockafellar (1968):

Proposition A.2 *If μ is finite and U and V are decomposable then $F(u)$ is a closed proper convex function, and its conjugate is*

$$F^*(v) = \int f^*\big(v(z)\big)\, d\mu(z).$$

Next proposition presents two additional results relating the properties of F and f:

Proposition A.3 *If μ is finite and U and V are decomposable then*

(i) *$v \in \partial F(u)$ if and only if $v(z) \in \partial f\big(u(z)\big)$, μ-a.e. over z.*

(ii) *If f is strictly convex, then so is F.*

Proof To show part (i), use first-order optimality for conjugates to obtain that $v \in \partial F(u)$ if and only if

$$F(u) + F^*(v) = \langle u, v \rangle \ . \tag{9}$$

Since Fenchel's inequality holds pointwise, i.e., $f\big(u(z)\big) + f^*\big(v(z)\big) \geq u(z)v(z)$, Eq. (9) is equivalent to

$$f\big(u(z)\big) + f^*\big(v(z)\big) = u(z)v(z), \quad \mu\text{-a.e. over } z,$$

which, again by first-order optimality for conjugates, is equivalent to

$$v(z) \in \partial f\big(u(z)\big), \quad \mu\text{-a.e. over } z,$$

completing the proof of part (i). Part (ii) can be shown by contradiction. Assume that F is not strictly convex, i.e., F is flat along a line segment connecting points u_1 and u_2 which differ on a set of non-zero measure. Let $u = (u_1 + u_2)/2$. The flatness of F means that $F(u) = [F(u_1) + F(u_2)]/2$, but pointwise, by convexity, $f\big(u(z)\big) \leq \big[f\big(u_1(z)\big) + f\big(u_2(z)\big)\big]/2$, so we must actually have $f\big(u(z)\big) = \big[f\big(u_1(z)\big) + f\big(u_2(z)\big)\big]/2$, μ-a.e. over z. Since u_1 and u_2 differ on a set of non-zero measure, we obtain that f cannot be strictly convex. ■

Fenchel's duality. Given pairings (X, Y) and (U, V) of Banach spaces and a continuous linear map $A : X \to U$, its *adjoint* is a linear map $A^\top : V \to Y$ defined by $\langle x, A^\top v \rangle = \langle Ax, v \rangle$. We finish this section by stating a version of Fenchel duality used in this paper. It is a rephrasing of the duality in Example 11' and Eq. (8.26) on page 50 of Rockafellar (1974), adapted to stronger conditions (specifically, $\operatorname{dom} F = X$ and $\operatorname{dom} G = U$):

Theorem A.4 *Let (X, Y) and (U, V) be pairings of Banach spaces. Let $F : X \to \mathbb{R}$ and $G : U \to \mathbb{R}$ be closed proper convex functions and $A : X \to U$ be a continuous linear operator. Then*

$$\inf_{x \in X} \Big[F(x) + G(Ax)\Big] = \max_{v \in V} \Big[-F^*(-A^\top v) - G^*(v)\Big] \ .$$

The point $\bar{x}$ is the primal minimizer if and only if there exists a dual maximizer $\bar{v}$ such that

$$-A^\top \bar{v} \in \partial F(\bar{x}) \ , \quad \bar{v} \in \partial G(A\bar{x}) \ .$$

Appendix B. Orlicz spaces

The duality result of Section 2 is an application of Fenchel's duality (Theorem A.4). As discussed in Section 2, the key challenge in applying the duality is the choice of appropriate pairings of Banach spaces. This appendix develops properties of specific Banach spaces, called *Orlicz spaces*, which will be sufficiently flexible to obtain pairings that satisfy our desiderata.

Orlicz spaces generalize $L_p(\mu)$ spaces introduced in Appendix A. The construction of an Orlicz space begins with a non-negative convex function $\theta : \mathbb{R} \to [0, \infty]$ symmetric around zero, not identical to zero, and with $\theta(0) = 0$, which serves the same role as the p-th power function in

the construction of $L_p(\mu)$. Given the function θ and a measure μ, we first define the unit ball of functions

$$\mathcal{B} := \left\{ f \text{ measurable} : \int \theta\big(f(z)\big) d\mu(z) \leq 1 \right\} ,$$

which is then used to define the norm $\|\cdot\|_\theta$:

$$\|f\|_\theta = \inf\{r \geq 0 : \ f \in r\mathcal{B}\}$$

where the norm equals ∞ if f is outside the scaled ball $r\mathcal{B}$ for all $r \geq 0$.

The *large Orlicz space* $L_\theta(\mu)$ and the *small Orlicz space* $M_\theta(\mu)$ are defined as

$$L_\theta(\mu) := \left\{ f \text{ measurable} : \ \exists r > 0, \int \theta(rf) d\mu < \infty \right\} ,$$
$$M_\theta(\mu) := \left\{ f \text{ measurable} : \ \forall r > 0, \int \theta(rf) d\mu < \infty \right\} .$$

From the definition it is clear that

$$L_\theta(\mu) = \{f : \ \|f\|_\theta < \infty\} ,$$

so for $p \geq 1$ and $\theta(s) = |s|^p$, we recover $L_p(\mu)$ spaces. The definition also implies $M_\theta(\mu) \subseteq L_\theta(\mu)$. The following proposition summarizes key properties of Orlicz spaces used in this paper. Parts (i–iv) are paraphrased from Proposition 1.4, Proposition 1.14, Proposition 1.18 and Theorem 2.2 of Léonard (2007):

Proposition B.1 *Let μ be a finite measure and $\theta : \mathbb{R} \to [0, \infty]$ be a closed convex function symmetric around zero, such that $\theta(0) = 0$ and neither θ nor its conjugate θ^* are identically zero. Then the following hold:*

(i) *θ^* is also symmetric around zero and $\theta^*(0) = 0$.*
(ii) *$L_\theta(\mu)$ and $M_\theta(\mu)$ are Banach spaces with the norm $\|\cdot\|_\theta$.*
(iii) *For all $f \in L_\theta(\mu)$ and $g \in L_{\theta^*}(\mu)$: $\int |fg| d\mu \leq 2\|f\|_\theta \|g\|_{\theta^*}$.*
(iv) *If θ is real-valued, i.e., $\operatorname{dom}\theta = \mathbb{R}$, then the topological dual of M_θ is isomorphic to L_{θ^*}.*
(v) *If θ is real-valued, i.e., $\operatorname{dom}\theta = \mathbb{R}$, then M_θ and L_{θ^*} are decomposable.*

Proof of (v) Let f be a bounded measurable function, say $|f| \leq a$. Then $\theta(rf(z)) \leq \theta(ra)$, so

$$\int \theta(rf) d\mu \leq \theta(ra)\mu(\mathcal{Z}) < \infty \text{ for all } r > 0,$$

implying $f \in M_\theta(\mu)$. Also, since $\operatorname{dom}\theta^* \neq \{0\}$ and $\theta^*(0) = 0$, there must be some $\varepsilon > 0$ such that $[-\varepsilon, \varepsilon] \subseteq \operatorname{dom}\theta^*$, and

$$\int \theta^*\left(\frac{\varepsilon}{a} f\right) d\mu \leq \theta^*(\varepsilon)\mu(\mathcal{Z}) < \infty$$

implying $f \in L_\theta^*(\mu)$. To argue that condition (ii) of Definition A.1 holds, note that if $f \in M_\theta(\mu)$ then any g with $|g| \leq |f|$ must also be in $M_\theta(\mu)$, and similarly for $L_{\theta^*}(\mu)$. ■

Appendix C. Rademacher complexity

This section collects various results from the literature on Rademacher complexity. To start, given a set of vectors $V \subseteq \mathbb{R}^n$, and letting $\sigma \in \{-1,+1\}^n$ denote a vector of n independent Rademacher random variables (i.e., $\Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = 1/2$ for all i), define the *Rademacher complexity* $\mathfrak{R}$ of V as

$$\mathfrak{R}(V) := \mathbb{E}\left(\sup_{v\in V} \frac{1}{n}\sum_{i=1}^n v_i\sigma_i\right).$$

To define the Rademacher complexity of a function f or function class $\mathcal{F}$ applied to a sample $\mathcal{S} := (z_i)_{i=1}^n$, define $f\circ\mathcal{S} := (f(z_i))_{i=1}^n \in \mathbb{R}^n$, and similarly overload $\mathcal{F}\circ\mathcal{S} \subseteq \mathbb{R}^n$, finally defining $\mathfrak{R}(\mathcal{F}) := \mathfrak{R}(\mathcal{F}\circ\mathcal{S})$. Note that these definitions match the presentation of *local Rademacher complexity* (Bartlett et al., 2005), whereas the original definition included an absolute value around the innermost summation (Bartlett and Mendelson, 2002; Boucheron et al., 2005).

The essential link between Rademacher complexity and deviation bounds is as follows.

Lemma C.1 (Shalev-Shwartz and Ben-David, 2014, Theorem 26.5) *Let loss ℓ and function class $\mathcal{F}$ be given. Then with probability at least $1-\delta$ over a draw of size n from μ,*

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left(\int \ell(f)d\mu - \int \ell(f)d\widehat{\mu}_n\right) &\leq 2\mathfrak{R}(\ell\circ\mathcal{F}\circ\mathcal{S}) + 4\sup_{\substack{z\in\mathcal{S}\\ f\in\mathcal{F}}}|\ell(f(z))|\sqrt{\frac{2\ln(4/\delta)}{n}}\\ &\leq 4\max\left\{1,\ \sup_{\substack{z\in\mathcal{S}\\ f\in\mathcal{F}}}|\ell(f(z))|\right\}\sqrt{\frac{\mathfrak{R}(\ell\circ\mathcal{F}\circ\mathcal{S})^2}{2} + \frac{4\ln(4/\delta)}{n}}.\end{aligned}$$

Thanks to Lemma C.1, the task of controlling deviations has been reduced to the task of approximating $\mathfrak{R}$. The following bounds are used throughout.

Lemma C.2 (See also Shalev-Shwartz and Ben-David, 2014, Chapter 26) *Let a collection of vectors $V \subseteq \mathbb{R}^n$ and a sample $\mathcal{S} := (z_i)_{i=1}^n$ be given.*

(i) *For any scalar $c \in R$ and any $v_0 \in \mathbb{R}^n$, $\mathfrak{R}(cV + v_0) \leq |c|\mathfrak{R}(V)$.*
(ii) *For sets $(V_j)_{j=1}^\infty$ with $V_j \subseteq \mathbb{R}^n$ and $0 \in V_j$ for all $j \geq 1$, it follows that $\mathfrak{R}(\cup_{j\geq 1}V_j) \leq \sum_{j\geq 1}\mathfrak{R}(V_j)$.*
(iii) *For $z_i \in \mathbb{R}^d$ and a set of linear predictors $\mathcal{W} := \{z \mapsto w\cdot z : w \in \mathbb{R}^d, \|w\|_1 \leq B\}$, it follows that $\mathfrak{R}(\mathcal{W}) = \mathfrak{R}(\mathcal{W}\circ\mathcal{S}) \leq B\sup_{z\in\mathcal{S}}\|z\|_\infty\sqrt{2\ln(2d)/n}$.*
(iv) *For any L-Lipschitz function $\ell : \mathbb{R}\to\mathbb{R}$, it follows that $\mathfrak{R}(\ell\circ V) \leq L\mathfrak{R}(V)$.*

Note that the aforementioned alternate form of $\mathfrak{R}$ using an absolute value breaks (i), whereas it strengthens (ii) by allowing the condition $0 \in V_j$ to be dropped.

Proof Proofs of parts (i), (iii), and (iv) can be found in (Shalev-Shwartz and Ben-David, 2014, Lemma 26.6, Lemma 26.11, Lemma 26.9); consequently, it only remains to handle (ii). For convenience, define $V_\infty := \cup_{j\geq 1}V_j$. Given any fixed $\sigma \in \{-1,+1\}^n$, the assumption $0 \in V_j$ implies

$$\sup_{v\in V_\infty} v\cdot\sigma \geq \sup_{v\in V_j} v\cdot\sigma \geq 0.$$

Table 1: Description of Datasets

Dataset	n (#examples)	s (average sparsity)	d (dimension)
20news	18845	93.9	101631
activity	165632	18.5	20
adult	48842	12.0	105
bio	145750	73.4	74
census	299284	32.0	401
covtype	581011	11.9	54
eeg	14980	14.0	14
ijcnn1	24995	13.0	22
kdda	8407751	36.3	19306083
kddcup2009	50000	58.4	71652
letter	20000	15.6	16
magic04	19020	10.0	10
maptaskcoref	158546	40.5	5944
mushroom	8124	22.0	117
nomao	34465	82.3	174
poker	946799	10.0	10
rcv1	781265	75.7	43001
shuttle	43500	7.0	9
skin	245057	2.9	3
vehv2binary	299254	48.6	105
w8a	49749	11.7	300

Consequently, by Tonelli's theorem,

$$\mathfrak{R}(V_\infty) = \mathbb{E}\left(\sup_{v\in V_\infty} \frac{1}{n} v\cdot\sigma\right) = \mathbb{E}\left(\sup_{j\geq 1}\sup_{v\in V_j} \frac{1}{n} v\cdot\sigma\right) \leq \mathbb{E}\left(\sum_{j\geq 1}\sup_{v\in V_j} \frac{1}{n} v\cdot\sigma\right)$$
$$= \sum_{j\geq 1}\mathbb{E}\left(\sup_{v\in V_j} \frac{1}{n} v\cdot\sigma\right) = \sum_{j\geq 1}\mathfrak{R}(V_j).$$

■

Appendix D. Experiments

In this appendix we demonstrate that the best performance on a wide variety of data sets can be obtained with little or no regularization. While there is some discussion of some methods' ability to seemingly avoid overfitting (Schapire et al., 1997; Friedman, 2000), this observation is primarily folklore, which served as a motivation for our experiments, depicted in Figure 3. They were conducted as follows:

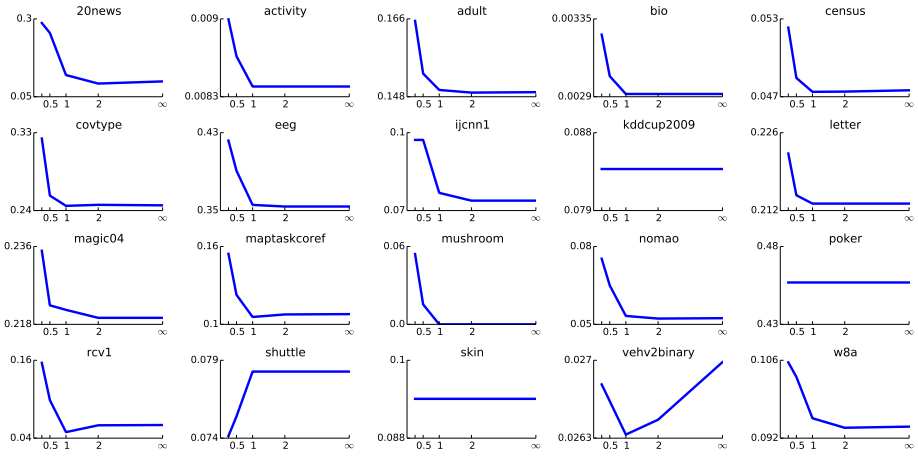

Figure 3: Proportion of classification errors on various testing sets of linear classifiers trained by applying L-BFGS to regularized logistic regression (ERM with logistic loss); test error is on the vertical axis, and exponent p of regularization coefficient $1/n^p$ is along the horizontal axis. For more detail, please see Appendix D.

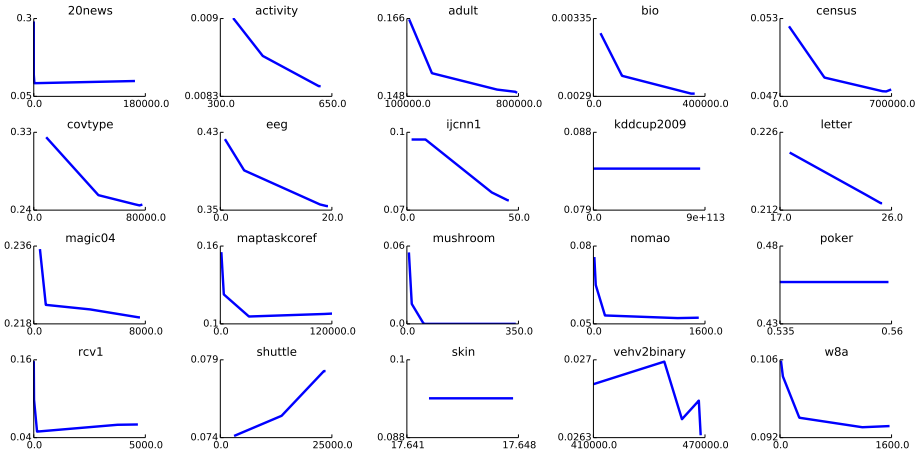

Figure 4: Companion plot to Figure 3; vertical axis is once again the proportion of classification errors, but the horizontal axis is now the quantity $\|w\|_2 \max_i \|x_i\|_2$, meaning the norm of the vector output by L-BFGS, scaled by the data norm. This quantity is relevant since it appears in the standard Rademacher bounds for linear functions (see Appendix C and Shalev-Shwartz and Ben-David, 2014, Chapter 26). For more detail, please see Appendix D.

1. We collected twenty datasets from a variety of sources (UCI, KDD Cup, libsvm data repository, and a few others), as described in Table 1.
2. Each dataset was split into 5 different (training, testing) pairs of size (80%, 20%).
3. ℓ was chosen to be the logistic loss $\ln(1+\exp(\cdot))$ and $\mathcal{H}$ consisted of the coordinates, yielding the setting of logistic regression.
4. We minimized the regularized empirical risk, i.e., $\hat{\mathcal{R}}_n(w) + \lambda\|w\|_2^2/2$, where λ was given the form $1/n^p$, where p ranged over $\{1/2, 1, 2, \infty\}$, with $1/n^\infty = 0$.
5. L-BFGS was applied to this regularized variant of $\widehat{\mathcal{R}}_n$ for each training/test split and each setting of the regularization parameter. Each point in Figure 3 is the median across the five splits of the data. Standard L-BFGS code was used (via `scikit-learn`), with very relaxed termination criteria in order to avoid early stopping (`pgtol` $= 10^{-9}$, `factr` $= 100$). In order to provide evidence that early stopping was avoided, please see Figure 4, which roughly captures the norms of the selected predictors.

Note that even as the norm of w increases, the classification error converges, and in most cases it is in fact minimized at large norms. It is essential that the plots depict classification error, whereby Theorem 1.2 and Proposition 1.3 explain why they behave stably. By contrast, if the goal were to recover specific iterates or control the loss itself, there are lower bounds indicating a dependence on norms is necessary (Levy et al., 2014).

Appendix E. Properties of classification losses

E.1. Basics

Lemma E.1

(i) *If* $\ell \in \mathbb{L}$, *then* $\lim_{z\to-\infty} \ell(z) = 0$, $\ell^*(0) = 0$, $\ell^*(s) = \infty$ *whenever* $s < 0$, *and* $\bar{s} \in \partial\ell(0)$ *satisfies* $\ell^*(\bar{s}) = \min_{s\in\mathbb{R}} \ell^*(s) = -\ell(0) < 0$.
(ii) *If* $\ell \in \mathbb{L}^{2+}$, *then* $\ell' > 0$ *and* $\lim_{z\to-\infty} \ell'(z) = 0$.
(iii) *If* $\ell \in \mathbb{L}^{2+}$, *then* $\liminf_{z\to-\infty} \ell''(z) = 0$.
(iv) *If* $\ell \in \mathbb{L}^{2+}$ *and* ℓ *is Lipschitz, then* $\liminf_{z\to\infty} \ell''(z) = 0$.
(v) *If* $\ell \in \mathbb{L}^{2+}$, *then* $\lim_{s\downarrow 0}(\ell^*)'(s) = -\infty$. *Additionally, if* $\lim_{r\to\infty} \ell'(r) =: L < \infty$, *then* $\lim_{s\uparrow L}(\ell^*)'(s) = \infty$.

Proof

(i) The first property follows from $\inf_{z\in\mathbb{R}} \ell(z) = 0$, which for convex ℓ with $\lim_{z\to-\infty} \ell(z) > 0$ implies ℓ is not nondecreasing.

The second property follows from $\ell^*(0) = \sup_{r\in\mathbb{R}}(0 \cdot r - \ell(r)) = -\inf_{s\in\mathbb{R}} \ell(s) = 0$.

Next, since $\lim_{z\to-\infty} \ell(z) = 0$, $s < 0$ implies

$$\ell^*(s) = \sup_{r\in\mathbb{R}}(rs - \ell(r)) \geq \lim_{r\to-\infty}(rs - \ell(r)) = \infty.$$

Lastly, because ℓ is closed, $\bar{s} \in \partial\ell(0)$ implies $0 \in \ell^*(\bar{s})$, which is the first order optimality condition, giving $\ell^*(\bar{s}) = \min_{z\in\mathbb{R}} \ell^*(z)$. Moreover, by Fenchel's inequality, $\ell^*(\bar{s}) + \ell(0) = 0 \cdot \bar{s} = 0$, meaning $\ell^*(\bar{s}) = -\ell(0)$, and lastly $\ell(0) > 0$, because $\ell \in \mathbb{L}$.

(ii) If there existed z' with $\ell'(z') = 0$, then $\ell'' > 0$ implies $\ell'(z'-1) < 0$, contradicting the fact that ℓ is nondecreasing.

Next, Mean Value Theorem grants for every $z < 0$ a $q_z \in [2z, z]$ such that

$$0 = \lim_{z\to-\infty} \ell(z) = \lim_{z\to-\infty} (\ell(2z) + \ell'(q_z)(z-2z)) = \lim_{z\to-\infty} (-z)\ell'(q_z),$$

which necessitates $\lim_{z\to-\infty} \ell'(z) = 0$ since ℓ' is nondecreasing and $\ell' \geq 0$.

(iii) Similarly to the above derivation for first derivatives, Mean Value Theorem grants for every $z < 0$ a $q_z \in [2z, z]$ such that

$$0 = \lim_{z\to-\infty} \ell'(z) = \lim_{z\to-\infty} (\ell'(2z) + \ell''(q_z)(z-2z)) = \lim_{z\to-\infty} (-z)\ell''(q_z),$$

which necessitates $\liminf_{z\to-\infty} \ell''(z) = 0$ by positivity of ℓ''.

(iv) Since $\lim_{z\to-\infty} \ell'(z) = 0$ (as above) and $\ell'' > 0$ and ℓ is Lipschitz, then there exists $L \geq 0$ with $\lim_{z\to\infty} \ell'(z) = L < \infty$. Similarly to the proof of the preceding property, Taylor's theorem grants for every $z > 0$ a $q_z \in [z, 2z]$ with

$$L = \lim_{z\to\infty} \ell'(z) = \lim_{z\to\infty} (\ell'(2z) + \ell''(q_z)(z-2z)) = L + \lim_{z\to\infty} (-z)\ell''(q_z),$$

which again necessitates $\liminf_{z\to\infty} \ell''(z) = 0$ by positivity of ℓ'.

(v) By strict convexity of ℓ, ℓ^* is differentiable over the interior of its domain (Hiriart-Urruty and Lemaréchal, 2001, Theorem E.4.1.1). By part (i), $\operatorname{dom} \ell^*$ includes 0 and $\bar{s} > 0$, so we can write $\lim_{s\downarrow 0} (\ell^*)'(s) = \lim_{s\downarrow 0} (\ell')^{-1}(s) = -\infty$. Where the last step follows because ℓ' is strictly increasing and $\lim_{r\to-\infty} \ell'(r) = 0$.

Given $\lim_{r\to\infty} \ell'(r) = L < \infty$, we obtain as before $\lim_{s\uparrow L} (\ell^*)'(s) = \lim_{s\uparrow 0} (\ell')^{-1}(s) = \infty$.

■

Proposition E.2 *Let $\ell \in \mathbb{L}^{2+}$ be given.*

(i) *The link ϕ is a monotone increasing bijection between $\mathbb{R}$ and $(0,1)$, and moreover continuously differentiable.*

(ii) *If ϕ is convex over $(-\infty, 0]$ and concave over $[0, \infty)$, then $L_\phi = \ell''(0)/(2\ell'(0))$. (This holds in particular for the logistic and exponential losses, which therefore have L_ϕ, respectively, equal to $1/4$ and $1/2$.)*

Proof

(i) Note that

$$\phi'(z) = \frac{\ell''(z)\ell'(-z) + \ell'(z)\ell''(-z)}{(\ell'(z) + \ell'(-z))^2}; \tag{10}$$

which is positive and continuous, because $\ell \in \mathbb{L}^{2+}$, so ϕ is increasing. Note that $\lim_{r\to\infty} \phi(r) = 1$ and $\lim_{r\to-\infty} \phi(r) = 0$, because ℓ' is increasing and $\lim_{r\to-\infty} \ell'(r) = 0$. The bijection statement follows by continuity.

(ii) By assumption, ϕ' is largest at 0. By the form of ϕ' given in Eq. (10) above, it follows that $\phi'(0) = \ell''(0)/(2\ell'(0))$. The convexity/concavity property may be manually checked for the exponential and logistic losses, since they respectively give ϕ'' to be

$$-\frac{4e^{2x}(e^{2x}-1)}{(1+e^{2x})^3} \qquad \text{and} \qquad -\frac{e^x(e^x-1)}{(1+e^x)^3}.$$

■

E.2. Elements of $\mathbb{L}_b^{2+}$

Lemma E.3 *Let finite non-null measure μ over $\mathcal{Z}$ and function $f : \mathcal{Z} \to \mathbb{R}$ be given with $f \geq 0$ μ-a.e. and $\int \exp(f)d\mu < \infty$. Set $b := \int \exp(f)d\mu/\mu(\mathcal{Z})$. Then $\int \exp(f/b)d\mu \leq \mu(\mathcal{Z})e^{1/e}$.*

Proof Note that $b \geq \int \exp(0)d\mu/\mu(\mathcal{Z}) = 1$. Consequently, the function $r \mapsto r^{1/b}$ is concave, and thus Jensen's inequality (applied to the normalized measure $\mu/\mu(\mathcal{Z})$) grants

$$\int \exp(f/b)d\mu = \mu(\mathcal{Z}) \int \exp(f)^{1/b} d\mu/\mu(\mathcal{Z}) \leq \mu(\mathcal{Z}) \left(\int \exp(f)d\mu/\mu(\mathcal{Z}) \right)^{1/b} = \mu(\mathcal{Z})b^{1/b}.$$

Next it will be shown that the function $g(z) := z^{1/z}$ is maximized over $(0, \infty)$ at $\bar{z} := e$, which gives the result. To this end, note

$$g'(z) = z^{1/z} z^{-2}(1 - \ln(z)),$$

which is positive for $z \in (0, \bar{z})$, zero at $\bar{z}$, and negative for $z > \bar{z}$. ■

Lemma E.4 *Let finite measure μ over $\mathcal{Z}$ and function $f : \mathcal{Z} \to \mathbb{R}$ be given with $f \geq 0$ μ-a.e. and $\mu(\mathcal{Z}) \leq 2$. If $\ell \in \mathbb{L}$ denotes the exponential loss $\ell = \exp$, then $\|f\|_\beta \leq \int \exp(f)d\mu/\mu(\mathcal{Z})$.*

Proof If $\int \exp(f)d\mu = \infty$, there is nothing to show, thus suppose $\int \exp(f)d\mu < \infty$. Since $\ell''(z) = \exp(z) \geq 1$ if $z \geq 0$ and ≤ 1 if $z \leq 0$, the Taylor expansion yields, for $z \geq 0$,

$$\exp(z) \geq 1 + z + z^2/2 \qquad \text{and} \qquad \exp(-z) \leq 1 - z + z^2/2.$$

Consequently, for any $z \geq 0$,

$$\exp(z) - (1+z) \geq z^2/2 \geq \exp(-z) - (1-z),$$

which means $\beta(z) = \exp(z) - (1+z)$ when $z \geq 0$. Combining this with Lemma E.3, setting $b := \int \exp(f)d\mu/\mu(\mathcal{Z})$ for convenience,

$$\int \beta(f/b)d\mu = \int \left(\exp(f/b) - 1 - f/b\right) d\mu \leq \mu(\mathcal{Z})(e^{1/e} - 1) \leq 1.$$

By the definition of $\|\cdot\|_\beta$, it follows that $\|f\|_\beta \leq b$. ■

Lemma E.5 *Let finite measure μ over $\mathcal{Z}$ and function $f : \mathcal{Z} \to \mathbb{R}$ be given with $f \geq 0$ μ-a.e.. If $\ell \in \mathbb{L}$ is L-Lipschitz, then $\|f\|_\beta \leq L \int \ell(f) d\mu / \ell'(0)$.*

Proof To start, for any $r \geq 0$, since ℓ is nondecreasing,

$$\begin{aligned}
\beta(r) &= \max\{\ell(r) - (\ell(0) + r\ell'(0)), \ell(-r) - (\ell(0) - r\ell'(0))\} \\
&\leq \max\{\ell(0) + rL - (\ell(0) + r\ell'(0)), \ell(0) - (\ell(0) - r\ell'(0))\} \\
&\leq r \max\{L - \ell'(0), \ell'(0)\}.
\end{aligned}$$

Setting $b := L \int \ell(f) d\mu / \ell'(0)$,

$$\begin{aligned}
\int \beta(f/b) d\mu &\leq \max\{L - \ell'(0), \ell'(0)\} \frac{\ell'(0) \int f d\mu}{L \int \ell(f) d\mu} \\
&\leq \max\{L - \ell'(0), \ell'(0)\} \frac{\ell'(0) \int f d\mu}{L(\ell(0) + \ell'(0) \int f d\mu)} \\
&\leq \max\{L - \ell'(0), \ell'(0)\} \frac{\ell'(0) \int f d\mu}{L\ell'(0) \int f d\mu} \\
&\leq 1,
\end{aligned}$$

where the last step follows because $\ell'(0) \leq L$. By the definition of $\|\cdot\|_\beta$, it follows that $\|f\|_\beta \leq b$. ■

Proposition E.6 *Let finite measure μ over $\mathcal{Z}$ with $\mu(\mathcal{Z}) \leq 2$ and hypotheses $\mathcal{H}$ be given. Then $\ell \in \mathbb{L}$ having a finite Lipschitz constant L entails $c_{\ell,\mu} \leq L/\ell'(0)$, and $\ell = \exp$ entails $c_{\ell,\mu} \leq 1/\mu(\mathcal{Z})$. Secondly, $\ell = \ln(1 + \exp(\cdot))$ entails $L_\phi = 1/4$, and $\ell = \exp$ entails $L_\phi = 1/2$. Thirdly, $\ell = \ln(1 + \exp(\cdot))$ entails $c_\ell = 2$, and $\ell = \exp$ entails $c_\ell = 1$. In particular, in either case, the loss is within $\mathbb{L}_{\mathrm{b}}^{2+}$.*

Proof Everything but the bounds on c_ℓ have already been provided by Lemma E.5, Lemma E.4, and Proposition E.2. For c_ℓ, the bound is immediate for $\ell = \exp$ (since then $\ell = \ell'$), thus consider $\ell = \ln(1 + \exp(\cdot))$. Noting the second-order Taylor expansion of $\ln$ along $[1, 1+q]$ with $q \leq 1$ is $\ln(1+q) \geq \ln(1) + q\ln'(1) + \inf_{s \in [1,2]} q^2 \ln''(s)/2$, then $r \leq 0$ implies

$$\ell(r) = \ln(1 + e^r) \geq e^r - \sup_{s \in [1,2]} \frac{e^{2r}}{2s^2} = e^r \left(1 - \frac{e^r}{2}\right) \geq \frac{e^r}{2} \geq \frac{e^r}{2(1+e^r)} = \frac{\ell'(r)}{2}.$$

■

Appendix F. Proof of Proposition 1.3

The proof of Proposition 1.3 is split into two lemmas; first, an upper bound establishing the general inequality, and second, an example showing the right-hand side of the inequality can be positive and tight. The proof of this upper bound is a straightforward consequence of standard manipulations for classification error (Devroye et al., 1996, Theorem 2.1).

Lemma F.1 *Let probability measure μ, hypotheses $\mathcal{H}$, and loss $\ell \in \mathbb{L}^{2+}$ be given. For any $w \in L_1(\mathcal{H})$,*

$$|\mathcal{R}_{\mathrm{z}}(Hw) - \mathcal{R}_{\mathrm{z}}(\bar{\eta}(\cdot,1) - 1/2)| \leq \int_{\bar{\eta}=1/2} (\eta_\mu(x,1) - \eta_\mu(x,-1))(1 - \mathbf{1}[\eta_w(x,1) \geq 1/2]) d\mu_{\mathcal{X}}(x) + \underbrace{2 \int_{\bar{\eta}\neq 1/2} \min\left\{1, \left(\frac{|\eta_\mu(x,1) - 1/2|}{|\bar{\eta}(x,1) - 1/2|}\right) |\bar{\eta}(x,1) - \eta_w(x,1)|\right\} d\mu_{\mathcal{X}}(x)}_{\star},$$

where $\star \to 0$ as $\int |\bar{\eta} - \eta_w| d\mu \to 0$.

Proof Following the derivation of Devroye et al. (1996, Theorem 2.1), for any $g : \mathcal{X} \to \{-1, +1\}$ and any $x \in \mathcal{X}$,

$$\begin{aligned}\Pr[g(X) \neq Y | X = x] &= 1 - \Pr[g(X) = Y | X = x] \\ &= 1 - (\mathbf{1}[g(x) = 1]\eta_\mu(x,1) + \mathbf{1}[g(x) = -1]\eta_\mu(x,-1)).\end{aligned}$$

Consequently, for any $g_1 : \mathcal{X} \to \{-1, +1\}$, $g_2 : \mathcal{X} \to \{-1, +1\}$, and any $x \in \mathcal{X}$,

$$\begin{aligned}&\Pr[g_1(X) \neq Y | X = x] - \Pr[g_2(X) \neq Y | X = x] \\ &= \eta_\mu(x,1)(\mathbf{1}[g_2(x) = 1] - \mathbf{1}[g_1(x) = 1]) + \eta_\mu(x,-1)(\mathbf{1}[g_2(x) = -1] - \mathbf{1}[g_1(x) = -1]) \\ &= (\eta_\mu(x,1) - \eta_\mu(x,-1))(\mathbf{1}[g_2(x) = 1] - \mathbf{1}[g_1(x) = 1]). \qquad (11)\end{aligned}$$

With this in mind, define $g_1(x) := \mathbf{1}[\eta_w(x,1) \geq 1/2]$ and $g_2(x) := \mathbf{1}[\bar{\eta}(x,1) \geq 1/2]$, whereby the signs of $(Hw)(x)$ and $\eta_w(x,1) - 1/2$ agree, and

$$\begin{aligned}&\mathcal{R}_{\mathrm{z}}(Hw) - \mathcal{R}_{\mathrm{z}}(\bar{\eta} - 1/2) \\ &\quad = \Pr[g_1(X) \neq Y] - \Pr[g_2(X) \neq Y] \\ &\quad = \underbrace{\int_{\bar{\eta}=1/2} (\Pr[g_1(X) \neq Y | X = x] - \Pr[g_2(X) \neq Y | X = x]) \, d\mu_{\mathcal{X}}(x)}_{\triangle} \\ &\qquad + \underbrace{\int_{\bar{\eta}\neq 1/2} (\Pr[g_1(X) \neq Y | X = x] - \Pr[g_2(X) \neq Y | X = x]) \, d\mu_{\mathcal{X}}(x)}_{\square}.\end{aligned}$$

To bound these terms, applying Eq. (11) to the first term and using $g_2(x) = 1$ along $\bar{\eta} = 1/2$ yields

$$\triangle = \int_{\bar{\eta}=1/2} (\eta_\mu(x,1) - \eta_\mu(x,-1))(1 - \mathbf{1}[g_1(x) = 1]) d\mu_{\mathcal{X}}(x).$$

For the second term, note

$$\mathbf{1}[g_1(x) \neq g_2(x)] \leq \min\left\{1, \frac{|\eta_w(x,1) - \bar{\eta}(x,1)|}{|\bar{\eta}(x,1) - 1/2|}\right\}.$$

Combining this with Eq. (11),

$$\begin{aligned}|\square| &\leq 2\int_{\bar\eta\neq 1/2} \min\left\{|\eta_\mu(x,1)-1/2|, \left(\frac{|\eta_\mu(x,1)-1/2|}{|\bar\eta(x,1)-1/2|}\right)|\eta_w(x,1)-\bar\eta(x,1)|\right\} d\mu_{\mathcal{X}}(x)\\ &\leq \star,\end{aligned}$$

with $\star$ given in the statement in the statement. To see that $\star \to 0$ as $\|\eta_w - \bar\eta\|_1 \to 0$, first note, for any $\sigma \in (0, 1/2]$, that

$$\begin{aligned}\star &\leq 2\int_{|\bar\eta-1/2|\in(0,\sigma)} 1 d\mu_{\mathcal{X}}(x) + 2\int_{|\bar\eta-1/2|>\sigma}\left(\frac{|\eta_\mu(x,1)-1/2|}{|\bar\eta(x,1)-1/2|}\right)|\eta_w(x,1)-\bar\eta(x,1)|d\mu_{\mathcal{X}}(x)\\ &\leq 2\int_{|\bar\eta-1/2|\in(0,\sigma)} 1 d\mu_{\mathcal{X}}(x) + \frac{1}{\sigma}\int |\eta_w(x,1)-\bar\eta(x,1)|d\mu_{\mathcal{X}}(x).\end{aligned}$$

Since the first term goes to 0 as $\sigma \to 0$, it suffices to choose $\sigma := \sqrt{\|\eta_w - \bar\eta\|_1}$ and the result follows. ■

In order to establish the tightness of the bound, consider any $\varepsilon \in [0, 1)$, let $\mathcal{X} = [-1, 1]$, and define the following probability measure μ over $\mathcal{X} \times \{-1, +1\} = \mathcal{Z}$:

$$\mu(x,\pm 1)\ \begin{cases} a := -1, & b := 1-\varepsilon;\\ \mu_{\mathcal{X}}(a) = \frac{1-\varepsilon}{2-\varepsilon}, & \mu_{\mathcal{X}}(b) = \frac{1}{2-\varepsilon};\\ \eta_\mu(a,+1) = 1, & \eta_\mu(b,+1) = 1.\end{cases}$$

Lemma F.2 *Let scalar $\varepsilon \in [0, 1)$, probability measure μ as above, hypotheses $\mathcal{H} := \{h\}$ where $h(x) = x$, and loss $\ell \in \mathbb{L}^{2+}$ be given. Then the sequence $(w_i)_{i=1}^\infty$ with $w_i := (-1)^i/i$ satisfies $\eta_{w_i} \to \bar\eta$ and*

$$\begin{aligned}\mathcal{R}_{\mathrm{z}}(Hw_i) - \mathcal{R}_{\mathrm{z}}(\bar\eta(\cdot,1)-1/2) &= \int_{\bar\eta=1/2}(\eta_\mu(x,1)-\eta_\mu(x,-1))\mathbf{1}[\eta_w(x,1)<1/2]d\mu_{\mathcal{X}}(x)\\ &= \begin{cases}\frac{1}{2-\varepsilon} & \text{when } i \text{ is odd},\\ \frac{1-\varepsilon}{2-\varepsilon} & \text{when } i \text{ is even}.\end{cases}\end{aligned}$$

Proof Note that $\mathcal{R}$ has primal optimum $\bar w = 0$: evaluating the gradient of $\mathcal{R}$ at $\bar w$ gives

$$-a\ell'(-a\bar w)\mu_{\mathcal{X}}(a) - b\ell'(-b\bar w)\mu_{\mathcal{X}}(b) = \ell'(0)\left(\frac{1-\varepsilon}{2-\varepsilon}\right) - \ell'(0)\left(\frac{1-\varepsilon}{2-\varepsilon}\right) = 0.$$

By Theorem 2.1, $\bar q = \ell'(0)$ μ-a.e., thus $\bar\eta = \phi(0) = 1/2$ μ-a.e., and $\mathcal{R}_{\mathrm{z}}(\bar\eta(\cdot,1)-1/2) = 0$.

Turning now to w_i, since $\bar\eta = 1/2$ and $\eta_\mu = \mathbf{1}[\bar\eta \geq 1/2]$ everywhere,

$$\begin{aligned}\mathcal{R}_{\mathrm{z}}(Hw_i) - \mathcal{R}_{\mathrm{z}}(\bar\eta(\cdot,1)-1/2) = \mathcal{R}_{\mathrm{z}}(Hw_i) &= \sum_{x\in\{a,b\}}\mu_{\mathcal{X}}(x)\mathbf{1}[\eta_{w_i}(x,1)<1/2]\\ &= \int_{\bar\eta=1/2}(\eta_\mu(x,1)-\eta_\mu(x,-1))\mathbf{1}[\eta_{w_i}(x,1)<1/2]d\mu_{\mathcal{X}}(x).\end{aligned}$$

Moreover, when i is odd, then $\mathcal{R}_z(Hw_i) = \mu_X(b)$, whereas i being even implies $\mathcal{R}_z(Hw_i) = \mu_X(a)$.

Lastly, the convergence statement follows since $w_i \to \bar{w}$, thus $\eta_{w_i} = \phi(Hw_i) \to \phi(H\bar{w}) = \bar{\eta}$ by continuity of ϕ (cf. Proposition E.2). ■

Proof (of Proposition 1.3) The proof follows by instantiating the bound in Lemma F.1 for each w_i, and applying $\limsup_{i\to\infty}$ to the absolute value of both sides. On the other hand, Lemma F.2 with any $\varepsilon \in [0, 1)$ provides the instance with $\star > 0$ and both $\limsup$s being equal. Note that the existence of oscilation exhibited in Lemma F.2 does not depend on our particular definition of $\operatorname{sign}(0)$. ■

Appendix G. Proofs from Section 2

To prove the main duality result (Theorem 2.1), we rely on a pairing of Orlicz spaces M_β and L_{β^*} implied by Proposition B.1.iv for a specific choice of β introduced in Eq. (6). We begin by showing how the norms $\|\cdot\|_\beta$ and $\|\cdot\|_{\beta^*}$ relate to the primal and dual objectives.

Recall that β is a symmetrized version of a loss $\ell \in \mathbb{L}$ with the first-order Taylor expansion at zero subtracted, and it thus represents the curvature of ℓ:

$$\beta(s) := \max\Big\{\ell(s) - \Big(\ell(0) + s\ell'(0)\Big),\ \ell(-s) - \Big(\ell(0) + (-s)\ell'(0)\Big)\Big\} \quad .$$

Note that this β satisfies the conditions on θ in Proposition B.1 and it is finite on $\mathbb{R}$, so we obtain the Banach space pairing between $M_\beta(\mu)$ with norm topology and $L_{\beta^*}(\mu)$ with weak* topology.

Lemma G.1 *Given a finite measure μ over $\mathcal{Z}$ and a loss function $\ell \in \mathbb{L}$, the following hold:*

(i) *If $f \in M_\beta(\mu)$, then $\int \ell(f)d\mu < \infty$.*

(ii) $\beta^*(s) \le \ell(0) + \min\big\{\ell^*\big(\ell'(0) - |s|\big),\ \ell^*\big(\ell'(0) + |s|\big)\big\}$.

(iii) *Let ν be any measure absolutely continuous with respect to μ, and let f denote its density with respect to μ, meaning $f := d\nu/d\mu$. Then $\int \ell^*(f)d\mu < \infty$ implies $f \in L_{\beta^*}(\mu)$.*

Proof

(i) Since $f \in M_\beta(\mu)$ means $\int \beta(f)d\mu < \infty$, the definition of β and property $\ell \ge 0$ grant

$$\begin{aligned}\int \ell(f)d\mu &\le \int \Big(\ell(f) + \ell(-f)\Big)d\mu \\ &= \int \Big(\ell(f) - \big(\ell(0) + \ell'(0)f\big) + \ell(-f) - \big(\ell(0) - \ell'(0)f\big)\Big)d\mu + 2\int \ell(0)d\mu \\ &\le 2\int \beta(f)d\mu + 2\ell(0)\mu(\mathcal{Z}) \\ &< \infty \ .\end{aligned}$$

(ii) For convenience, define

$$\ell_+(r) := \ell(r) - \big(\ell(0) + r\ell'(0)\big) \qquad \text{and} \qquad \ell_-(r) := \ell(-r) - \big(\ell(0) - r\ell'(0)\big) \ ,$$

and note (e.g., from definition of conjugate or by Theorem 12.3 of Rockafellar, 1970) that

$$\ell_+^*(s) := \ell(0) + \ell^*\big(\ell'(0) + s\big) \qquad \text{and} \qquad \ell_-^*(s) := \ell(0) + \ell^*\big(\ell'(0) - s\big) \ .$$

Since $\beta = \max\{\ell_+, \ell_-\}$, then, by definition of conjugate, $\beta^* \leq \min\{\ell_+^*, \ell_-^*\}$, yielding the result.

(iii) Let $\bar{s} := \ell'(0)$. By Lemma E.1.i, ℓ^* is minimized at $\bar{s} > 0$, so it must be non-increasing on $[0, \bar{s}]$ and non-decreasing on $[\bar{s}, \infty)$. Also, by Lemma E.1.i, $\ell^*(0) = 0$, so $\ell^* \leq 0$ on $[0, \bar{s}]$. Part (ii) therefore implies

$$\beta^*(s) \leq \begin{cases} \ell(0) & \text{if } |s| \leq \bar{s} \\ \ell(0) + \ell^*(2|s|) & \text{if } |s| > \bar{s}. \end{cases}$$

Let $f = d\nu/d\mu$, i.e., $f = |f|$ (μ-a.e.) and assume that $\int \ell^*(f) d\mu < \infty$. Using the previous bound on β^*, write

$$\begin{aligned} \int \beta^*(f/2) d\mu &\leq \ell(0)\mu(\mathcal{Z}) + \int_{f/2 > \bar{s}} \ell^*(f) d\mu \\ &= \ell(0)\mu(\mathcal{Z}) + \int \ell^*(f) d\mu - \int_{f/2 \leq \bar{s}} \ell^*(f) d\mu \\ &\leq \ell(0)\mu(\mathcal{Z}) + \int \ell^*(f) d\mu + \ell(0)\mu\left(\{f/2 \leq \bar{s}\}\right) \\ &< \infty \ , \end{aligned}$$

where the next to last step follows, because $\ell^*(s) \geq \ell^*(\bar{s}) = -\ell(0)$ by Lemma E.1.i.

■

Proof (of Theorem 2.1) The duality law will be proved via Fenchel's duality (Theorem A.4). To begin, we need to define Banach space pairings. One of them is $\big(L_1(\mathcal{H}),\, L'\big)$ where L' is the topological dual of $L_1(\mathcal{H})$ and the other is $\big(M_\beta(\mu),\, L_{\beta^*}(\mu)\big)$, which is a valid pairing as argued at the beginning of this appendix.

We invoke Theorem A.4 with $F : L_1(\mathcal{H}) \to \mathbb{R}$, $G : M_\beta(\mu) \to \mathbb{R}$ defined by

$$F(w) = 0 \text{ for all } w, \quad G(f) = \int \ell(f) d\mu$$

and $A : L_1(\mathcal{H}) \to M_\beta(\mu)$ defined as in Section 1. Note that $F^*(u) = \mathbb{I}[u = 0]$ where $\mathbb{I}$ denotes the convex indicator, yielding the constraint $A^\top q = 0$. To prove Eq. (7), it remains to show that A is continuous as a map from $L_1(\mathcal{H})$ to $M_\beta(\mu)$, G is finite on $M_\beta(\mu)$ and

$$G^*(q) = \int \ell^*(q) d\mu \ .$$

Finiteness of G follows by Lemma G.1.i; the expression for the conjugate G^* follows by Proposition A.2, because $M_\beta(\mu)$ and $L_{\beta^*}(\mu)$ are decomposable (by Proposition B.1). Finally, to argue

continuity of A, consider $w, w' \in L_1(\mathcal{H})$. From the definition of A, $|(Aw)(z)| \leq \|w\|_1$, so Aw is a bounded measurable function and hence in $M_\beta(\mu)$ (by decomposability). Also,

$$\big|\big(A(w'-w)\big)(z)\big| \leq \|w'-w\|_1 \ . \tag{12}$$

Let $f_1(z) = 1$ for all z. For any f and g such that $|f| \leq |g|$, we have $\|f\|_\beta \leq \|g\|_\beta$, so Eq. (12) implies

$$\big\|A(w'-w)\big\|_\beta \leq \|w'-w\|_1 \|f_1\|_\beta \ ,$$

showing the continuity of A, because $\|f_1\|_\beta$ is finite (by decomposability).

It remains to show the properties of the dual optima:

(i) The bound follows since $\ell^*(s) = \infty$ whenever $s < 0$ by Lemma E.1.

(ii) Any dual optimum $\bar{q}$ may be modified on a μ-null set to obtain $\hat{q}$ satisfying the condition. To start, define $S := \{x \in \mathcal{X} : \bar{q}(x,1) = \bar{q}(x,-1) = 0\}$; from part (i), $\bar{q} \geq 0$ (μ-a.e.), so it suffices to produce $\hat{q}$ by modifying $\bar{q}$ on a μ-null subset of S.

Recall that $\eta_\mu(x,y)$ represents the conditional probability of y given x, i.e., $d\mu(x,y) = \eta_\mu(x,y) d\mu_{\mathcal{X}}(x)$ and $\eta_\mu(x,-1) + \eta_\mu(x,1) = 1$. We will write $\mathcal{Y} = \{-1,1\}$. First consider those points where $\eta_\mu(x,y) \in (0,1)$; in particular, the set

$$S_0 := \{x \in S : \eta_\mu(x,1) \in (0,1)\} \, ,$$

and, for the sake of contradiction, suppose that $\mu_{\mathcal{X}}(S_0) > 0$. Pick $\bar{s} \in \partial\ell(0)$, whereby $\ell^*(\bar{s}) < \ell^*(0) = 0$ by Lemma E.1. Define $q \in L_{\beta^*}(\mu)$ as

$$q(x,y) := \begin{cases} \bar{q}(x,y) & \text{when } x \notin S_0, \\ \bar{s} & \text{when } x \in S_0 \text{ and } \eta_\mu(x,-y) \geq \eta_\mu(x,y), \\ \bar{s} \cdot \frac{\eta_\mu(x,-y)}{\eta_\mu(x,y)} & \text{when } x \in S_0 \text{ and } \eta_\mu(x,-y) < \eta_\mu(x,y). \end{cases}$$

We show that q is dual-feasible and achieves a better objective value than $\bar{q}$. By construction, $q \in L_{\beta^*}(\mu)$ (since $\bar{q} \in L_{\beta^*}$, which is decomposable, and the adjustment is bounded), and moreover, for every $w \in L_1(\mathcal{H})$,

$$\begin{aligned} (A^\top q)(w) &= \int_{S_0 \times \mathcal{Y}} (Aw) q \, d\mu + \int_{S_0^c \times \mathcal{Y}} (Aw) \bar{q} \, d\mu \\ &= \int_{S_0} (Hw)(x) \Big(q(x,-1)\eta_\mu(x,-1) - q(x,1)\eta_\mu(x,1) \Big) d\mu_{\mathcal{X}}(x) \\ &\quad + \left(\int (Aw) \bar{q} \, d\mu - \int_{S_0 \times \mathcal{Y}} (Aw) \bar{q} \, d\mu \right) \\ &= 0 + (0 - 0) \ , \end{aligned}$$

where the last step follows from the definition of q, feasibility of $\bar{q}$ and the fact that $S_0 \subseteq S$. Thus, q is feasible. On the other hand,

$$\int \ell^*(q) d\mu = \int \ell^*(\bar{q}) d\mu + \int_{S_0 \times \mathcal{Y}} \ell^*(q) d\mu \ ,$$

because $\bar{q} = 0$ along $S_0 \times \mathcal{Y}$. By construction, $q \in (0, \bar{s}]$ along $S_0 \times \mathcal{Y}$. Further, $\ell^*(s) < 0$ for $s \in (0, \bar{s}]$ by Lemma E.1, so $\ell^*(q) < 0$ along $S_0 \times \mathcal{Y}$. Hence, $\mu_{\mathcal{X}}(S_0) > 0$ implies q attains a lower objective value than $\bar{q}$, a contradiction; thus $\mu_{\mathcal{X}}(S_0) = 0$.

It has been shown that $\bar{q}(x, y) + \bar{q}(x, -y) > 0$ over (x, y) with $\eta_\mu(x, y) \in (0, 1)$, μ-a.e.; consequently, it suffices to consider (x, y) with $\eta_\mu(x, y) \in \{0, 1\}$. Define $\hat{q} \in L_{\beta^*}(\mu)$ as

$$\hat{q}(x, y) := \begin{cases} \bar{q}(x, y) & \text{when } \eta_\mu(x, y) \in (0, 1], \\ \bar{s} & \text{when } \eta_\mu(x, y) = 0. \end{cases}$$

Since the adjustment is only on points where $\eta_\mu(x, y) = 0$, then $\hat{q} = \bar{q}$ μ-a.e., and thus is also a dual solution. Furthermore, since $\mu_{\mathcal{X}}(S_0) = 0$, then $\mu_{\mathcal{X}}$-a.e. over $x \in S$, we have $\hat{q}(x, -1) + \hat{q}(x, 1) \geq \bar{s} > 0$ as desired.

(iii) This follows directly from Theorem A.4 and Proposition A.3.i.

(iv) Consider a sequence $(w_i)_{i=1}^{\infty}$ minimizing the primal. By Eq. (7) and since $A^\top \bar{q} = 0$, this means that

$$\int \ell(Aw_i) d\mu + \int \ell^*(\bar{q}) d\mu - \langle A^\top \bar{q},\, w_i \rangle \;\to\; 0 \tag{13}$$

as $i \to \infty$. Let $r_i = Aw_i$. Since $\langle A^\top \bar{q},\, w_i \rangle = \langle \bar{q},\, Aw_i \rangle = \langle \bar{q}, r_i \rangle$, Eq. (13) can be rearranged to

$$\int \Big[\ell(r_i) + \ell^*(\bar{q}) - \bar{q} r_i \Big] d\mu \;\to\; 0 \;.$$

By Fenchel's inequality, the integrand is non-negative, so we actually have

$$\ell(r_i(z)) + \ell^*(\bar{q}(z)) - \bar{q}(z) r_i(z) \;\to\; 0 \quad \mu\text{-a.e. over } z \in \mathcal{Z}. \tag{14}$$

Denote the set of points z where ℓ^* is differentiable at $\bar{q}(z)$ as S. Define $\bar{r}(z) := (\ell^*)'(\bar{q}(z))$ for $z \in S$. Over $z \in S$, we have by first-order optimality for conjugates that $\ell^*(\bar{q}) = \bar{q}\bar{r} - \ell(\bar{r})$, and $\bar{q} = \ell'(\bar{r})$, and thus Eq. (14) implies

$$\ell(r_i) - \ell(\bar{r}) - \ell'(\bar{r})(r_i - \bar{r}) \;\to\; 0 \quad \mu\text{-a.e. over } z \in S.$$

Hence, from strict convexity of ℓ we obtain that $r_i \to \bar{r}$, μ-a.e. over $z \in S$. Now, let $S_{\mathcal{X}} := \{x \in \mathcal{X} : (x, 1) \in S \text{ and } (x, -1) \in S\}$ be the set of points x where ℓ^* is differentiable at both $\bar{q}(x, 1)$ and $\bar{q}(x, -1)$. From the definition of r_i, we have $r_i(x, 1) + r_i(x, -1) = 0$ and thus we must also have $\bar{r}(x, 1) + \bar{r}(x, -1) = 0$, $\mu_{\mathcal{X}}$-a.e. over $x \in S_{\mathcal{X}}$. Unrolling the definition of $\bar{r}$ yields the desired result.

(v) If ℓ is differentiable, then ℓ^* is strictly convex (Hiriart-Urruty and Lemaréchal, 2001, Theorem E.4.1.2), whereby $\int \ell^* d\mu$ is also strictly convex by Proposition A.3.ii, and thus the dual optimizer is unique up to μ-null sets.

■

To close, note an additional technical property of $\bar{q}$ which will be useful in various proofs.

Lemma G.2 *Given finite measure μ, hypotheses $\mathcal{H}$, and loss $\ell \in \mathbb{L}^{2+}$ with $L := \lim_{r\to\infty} \ell'(r)$, it follows that every dual optimum $\bar{q}$ satisfies $\mu(\{z \in \mathcal{Z} : \bar{q}(z) \geq L\}) = 0$.*

Proof Note that ℓ^* is strictly convex (by differentiability of ℓ) and differentiable everywhere except possibly at the endpoints of its domain (by strict convexity of ℓ). If $L = \infty$, there is nothing to show, thus suppose $L < \infty$, which entails $\operatorname{dom}(\ell^*) \subseteq [0, L]$ (since the image of the derivative map ℓ' is the domain of the conjugate derivative map $(\ell^*)'$, and this coincides, up to the endpoints, with $\operatorname{dom} \ell^*$). So it suffices to show that $\mu(\bar{q} = L) = 0$.

Note that $\ell'(0) \in (0, L)$ since $\ell'' > 0$. Define a scalar $N := (L + \ell'(0))/2$, set $D := \{z \in \mathcal{Z} : \bar{q}(z) \in (0, L]\}$, and partition D into the three pieces

$$R_1 := \{z \in \mathcal{Z} : \bar{q}(z) \in (0, N]\},$$
$$R_2 := \{z \in \mathcal{Z} : \bar{q}(z) \in (N, L)\},$$
$$R_3 := \{z \in \mathcal{Z} : \bar{q}(z) = L\}.$$

We next study the integral $\int \ell^*((1-\alpha)\bar{q})d\mu$ for small values of α over these pieces.

(R_2) Since ℓ^* is increasing along $[\ell'(0), L]$, then every sufficiently small $\alpha > 0$ and every $z \in R_2$ satisfies $\ell^*((1-\alpha)\bar{q}(z)) < \ell^*(\bar{q}(z))$, and in particular

$$\int_{R_2} \ell^*((1-\alpha)\bar{q})d\mu \leq \int_{R_2} \ell^*(\bar{q})d\mu.$$

(R_1) Consider the function

$$F(\alpha) = \int_{R_1} \ell^*((1-\alpha)\bar{q})d\mu.$$

This is a univariate convex function which is finite on a neighborhood of 0. Pick $\tau > 0$ such that $[-\tau, \tau]$ lies in this neighborhood. Since this is a closed bounded subset of the relative interior of $\operatorname{dom} F$, we obtain (by Rockafellar, 1970, Theorem 10.4) that F is Lipschitz-continuous on $[-\tau, \tau]$. Let L' be its Lipschitz constant on $[-\tau, \tau]$. For $|\alpha| \leq \tau$, we obtain

$$\int_{R_1} \ell^*((1-\alpha)\bar{q})d\mu \leq \alpha L' + \int_{R_1} \ell^*(\bar{q})d\mu.$$

(R_3) Note $\lim_{z\uparrow L}(\ell^*)'(z) = \infty$ (by Lemma E.1), thus the definition of subgradient grants

$$\begin{aligned}\int_{R_3} \ell^*((1-\alpha)\bar{q})d\mu &= \mu(R_3)\ell^*((1-\alpha)L) \\ &\leq \mu(R_3)\left(\ell^*(L) - (\ell^*)'((1-\alpha)L)(L-(1-\alpha)L)\right) \\ &= -\alpha L\mu(R_3)(\ell^*)'((1-\alpha)L) + \int_{R_3} \ell^*(\bar{q})d\mu.\end{aligned}$$

To finish, first note $\bar{q} \in [0, L]$ for μ-a.e. $z \in \mathcal{Z}$ (since otherwise $\int \ell^*(\bar{q})d\mu > \int \ell^*(0)d\mu = 0$), and $\ell^*((1-\alpha)\bar{q}) = 0$ wherever $\bar{q} = 0$. Combining these pieces, since $\bar{q}$ is optimal and $(1-\alpha)\bar{q}$ is

feasible for $\alpha \in [0,1]$, then for sufficiently small $\alpha > 0$,

$$\begin{aligned}
\int \ell^*(\bar{q})d\mu &\leq \int \ell^*((1-\alpha)\bar{q})d\mu \\
&= \int_{R_1} \ell^*((1-\alpha)\bar{q})d\mu + \int_{R_2} \ell^*((1-\alpha)\bar{q})d\mu + \int_{R_3} \ell^*((1-\alpha)\bar{q})d\mu \\
&\leq \alpha L' + \int_{R_1} \ell^*(\bar{q})d\mu + \int_{R_2} \ell^*(\bar{q})d\mu - \alpha L\mu(R_3)(\ell^*)'((1-\alpha)L) + \int_{R_3} \ell^*(\bar{q})d\mu. \\
&= \alpha\Big(L' - L\mu(R_3)(\ell^*)'((1-\alpha)L)\Big) + \int \ell^*(\bar{q})d\mu,
\end{aligned}$$

which rearranges to give

$$L\mu(R_3)\underbrace{(\ell^*)'((1-\alpha)L)}_{\triangle} \leq L'.$$

Since $L > 0$ and $\triangle \to \infty$ as $\alpha \downarrow 0$ whereas L' is constant, it follows that $\mu(R_3) = 0$. ■

Appendix H. Proof of Lemma 3.2 and Corollary 3.3

This brief appendix section collects proofs of two results from the introductory part of Section 3.

Proof (of Lemma 3.2) Applying Theorem 2.1, to both μ and μ_D,

$$\begin{aligned}
\inf_{w \in L_1(\mu)} \mathcal{R}(w) &= \max_{q \in L_{\beta^*}(\mu):\, A^\top q = 0} \left[-\int \ell^*(q)d\mu \right], \\
\inf_{w \in L_1(\mu)} \mathcal{R}(w;\mu_D) &= \max_{q \in L_{\beta^*}(\mu_D):\, A^\top q = 0} \left[-\int_D \ell^*(q)d\mu \right].
\end{aligned}$$

Of course, $\bar{q}$ attains the first dual maximum over μ; note, as follows, that it also attains the dual maximum over μ_D. First, $\bar{q}$ is feasible for the second problem, since $\bar{q} \in L_{\beta^*}(\mu)$ and $\bar{q} = 0$ on D^c, so we also have $\bar{q} \in L_{\beta^*}(\mu_D)$, and for every $v \in L_1(\mathcal{H})$,

$$0 = \int (Av)\bar{q}\, d\mu = \int_D (Av)\bar{q}\, d\mu.$$

Furthermore, since $\ell \in \mathbb{L}$ implies $\ell^*(0) = 0$ (by Lemma E.1), it follows that

$$\int \ell^*(\bar{q})d\mu = \int_D \ell^*(\bar{q})d\mu.$$

Consequently,

$$\max_{\substack{q \in L_{\beta^*}(\mu) \\ A^\top q = 0}} \left[-\int \ell^*(q)d\mu \right] = -\int \ell^*(\bar{q})d\mu = -\int_D \ell^*(\bar{q})d\mu \leq \max_{\substack{q \in L_{\beta^*}(\mu_D) \\ A^\top q = 0}} \left[-\int_D \ell^*(q)d\mu \right]. \quad (15)$$

Now consider any dual optimum $\bar{q}_D$ over μ_D, and set $\hat{q}(z) := \bar{q}_D(z)\mathbf{1}[z \in D]$. Mimicking the derivations above, $\hat{q}$ is feasible and optimal over D (indeed, $\bar{q}_D$ and $\hat{q}$ only differ on a μ_D-null set).

Similarly, however, $\hat{q}$ is also feasible for the full problem over μ_D, and $\int \ell^*(\hat{q})\mu_D = \int \ell^*(\hat{q})d\mu$, implying that the inequality in Eq. (15) is an equality, and $\bar{q}$ and $\hat{q}$ are optimal for both μ and μ_D. ■

Proof (of Corollary 3.3) Using the fact that the dual and thus also primal optimal values coincide for μ and μ_D, as well as the fact that $\ell \geq 0$, we obtain

$$\mathcal{E}(w;\mu_D) = \int_D \ell(Aw)d\mu - \inf_{v\in L_1(\mathcal{H})} \int_D \ell(Av)d\mu \leq \int \ell(Aw)d\mu - \inf_{v\in L_1(\mathcal{H})} \int \ell(Av)d\mu = \mathcal{E}(w)$$

directly, and similarly

$$\begin{aligned}
\mathcal{R}(w;\mu_{D^c}) &= \int \ell(Aw)d\mu - \int_D \ell(Aw)d\mu \\
&\leq \int \ell(Aw)d\mu - \inf_{v\in L_1(\mathcal{H})} \int_D \ell(Av)d\mu \\
&= \int \ell(Aw)d\mu - \inf_{v\in L_1(\mathcal{H})} \int \ell(Av)d\mu \\
&= \mathcal{E}(w).
\end{aligned}$$

■

Appendix I. Proofs from Section 3.1

Before proving Lemma 3.5 in full, we establish a general form of its first part. Unlike the proof of the second part of Lemma 3.5, the first part does not rely upon the structure of $\mathcal{D}^c$ in any way; indeed it is simply Markov's inequality.

Lemma I.1 *Let finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, and arbitrary set $C \subseteq \mathcal{Z}$ be given. Then for any $w \in L_1(\mathcal{H})$ and $r > 0$, the set $S_r := \{z \in C : \ell\big((Aw)(z)\big) \geq r\}$ satisfies $\mu(S_r) \leq \mathcal{R}(w;\mu_C)/r$.*

Proof Emulating the proof of Markov's inequality, every $z \in \mathcal{Z}$ satisfies

$$r\mathbf{1}[z \in S_r] \leq r\mathbf{1}[\ell\big((Aw)(z)\big) \geq r] \leq \ell\big((Aw)(z)\big),$$

thus integrating both sides along C and dividing by r gives

$$\mu(S_r) \leq \frac{\int_C \ell(Aw)d\mu}{r} = \frac{\mathcal{R}(w;\mu_C)}{r}.$$

■

Proof (of Lemma 3.5) Part (i) is proved by applying Lemma I.1 with $C := D^c$, and then applying Corollary 3.3 for the inequality $\mathcal{R}(w;\mu_{D^c}) \leq \mathcal{E}(w)$.

For part (ii), first note that if $r \geq \ell(0)$, we are done, because $|\bar{\eta} - \eta_w| \leq 1$. Now consider $r < \ell(0)$. Since $\bar{\eta} = 1$ for μ-a.e. $z \in D^c$ and $\eta_w \geq 0$ by definition, then

$$\int_{D^c \setminus S_r} |\bar{\eta} - \eta_w| \, d\mu = \int_{D^c \setminus S_r} (1 - \eta_w) d\mu = \int_{D^c \setminus S_r} \frac{\ell'((Aw)(x,y))}{\ell'((Aw)(x,y)) + \ell'((Aw)(x,-y))} d\mu =: \heartsuit,$$

thus it remains to control $\heartsuit$. Since every $z \in D^c \setminus S_r$ has $\ell\big((Aw)(z)\big) < r < \ell(0)$, the increasing property of ℓ implies $(Aw)(z) \leq 0$. Consequently, it follows that $\ell'\big((Aw)(z)\big) \leq c_\ell \ell\big((Aw)(z)\big) < c_\ell r$, and also that $\ell'((Aw)(x,-y)) = \ell'(-(Aw)(x,y)) \geq \ell'(0)$ since ℓ' is nondecreasing by convexity. Combining these bounds,

$$\heartsuit = \int_{D^c \setminus S_r} \frac{1}{1 + \ell'((Aw)(x,-y))/\ell'((Aw)(x,y))} d\mu(x,y) \leq \frac{\mu(D^c \setminus S_r)}{1 + \ell'(0)/(c_\ell r)},$$

which gives the desired bound after rearrangement, noting that $c_\ell r > 0$. ■

In order to prove Lemma 3.6, it will be necessary to establish an additional structural property of dual optima. In particular, recall the function $\bar{f}$, which is used in the proof of Lemma 3.6, and which is equal to $(\ell')^{-1}(\bar{q})$ whenever $(\ell')^{-1}$ is defined for both $\bar{q}(x,y)$ and $\bar{q}(x,-y)$. It is this final condition—needing both (x,y) and $(x,-y)$—which requires the extra work here.

For the purposes of Lemma 3.6, it will suffice to establish that μ-a.e $(x,y) \in \mathcal{D}$ satisfies $(x,-y) \in \mathcal{D}$, which is precisely the following lemma. This result is in fact a consequence of Lemma 3.5: the idea is that for those points with $(x,y) \in \mathcal{D}$ but $(x,-y) \in \mathcal{D}^c$, applying Lemma 3.5 grants that every low error predictor must achieve small error on this latter set. But this leads to a contradiction, since it necessitates that the error on the mirrored points, which reside in $\mathcal{D}$, must be large.

Lemma I.2 *Let finite measure μ, hypotheses $\mathcal{H}$, and loss $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$ be given. Then there exists a dual optimum $\bar{q}$ and corresponding difficult set $\mathcal{D}$ such that μ-a.e. over $(x,y) \in \mathcal{D}$ we also have $(x,-y) \in \mathcal{D}$.*

Proof Let an arbitrary dual optimum $\bar{q}_0$ be given as provided by Theorem 2.1, and let $\mathcal{D}_0$ denote the corresponding difficult set. If this provided $\bar{q}_0$ already satisfies the necessary properties, the proof is done, therefore suppose it does not.

Define three sets

$$\begin{aligned}
K_0 &:= \{(x,y) \in \mathcal{D}_0 : (x,-y) \in \mathcal{D}_0^c, \eta_\mu(x,y) = 0\}, \\
K_1 &:= \{(x,y) \in \mathcal{D}_0 : (x,-y) \in \mathcal{D}_0^c, \eta_\mu(x,y) = 1\}, \\
K_+ &:= \{(x,y) \in \mathcal{D}_0 : (x,-y) \in \mathcal{D}_0^c, \eta_\mu(x,y) \in (0,1)\},
\end{aligned}$$

and an adjusted dual optimum

$$\bar{q}(x,y) := \begin{cases} 0 & \text{when } (x,y) \in K_0, \\ \ell'(0) & \text{when } (x,-y) \in K_1, \\ \bar{q}_0(x,y) & \text{otherwise.} \end{cases}$$

Since $\mu(K_0) = 0 = \mu(\{(x,y) : (x,-y) \in K_1\})$ by construction, then $\bar{q} = \bar{q}_0$ μ-a.e., meaning $\bar{q}$ is also a dual optimum to Eq. (7). Defining $\mathcal{D} := \{z \in \mathcal{Z} : \bar{q}(z) > 0\}$, if $(x,y) \in \mathcal{D}$ and $(x,-y) \in \mathcal{D}^c$, then it must hold that $(x,y) \in K_+$. The proof is done if $\mu(K_+) = 0$; this will constitute the remainder of the proof.

Assume contradictorily that $\mu(K_+) > 0$. Define

$$U_\xi := \{(x,y) \in K_+ : \min\{\eta_\mu(x,y), \eta_\mu(x,-y)\} \geq \xi\}.$$

By continuity of measures, $\lim_{\xi \downarrow 0} \mu(U_\xi) = \mu(\cup_{\xi>0} U_\xi) = \mu(K_+)$, thus there exists a fixed $\tau > 0$ so that $U := U_\tau$ has $\mu(U) \geq \tau$. For convenience, define $U_- := \{(x,-y) : (x,y) \in U\}$ (and use S_- for this "flipped sign" transformation of any set $S \subseteq \mathcal{Z}$). By the conditions on U, then $\mu(U_-) \geq \tau\mu(U) \geq \tau^2$, and for any set $C \subseteq \mathcal{Z}$,

$$\mu(U \cap C_-) \geq \tau\mu(U_- \cap C). \tag{16}$$

Now choose $\varepsilon_0 > 0$ so that $\ell(-\ell^{-1}(\sqrt{\varepsilon_0})) > 6\mathcal{R}(0)/\tau^3$, set $\varepsilon := \min\{\varepsilon_0, \tau^4/4, \mathcal{R}(0)\}$, and choose $w \in L_1(\mathcal{H})$ with $\mathcal{E}(w) \leq \varepsilon$. Applying Lemma 3.5 to w with $r := \sqrt{\varepsilon}$, the set

$$S_r := \{z \in \mathcal{D}^c : \ell((Aw)(z)) \geq r\}$$

satisfies $\mu(S_r) \leq \varepsilon/r = r$. For convenience, define $V := \mathcal{D}^c \setminus S_r$, whereby $\mu(V) \geq \mu(\mathcal{D}) - r$, and every $z \in V$ has $\ell((Aw)(z)) \leq \sqrt{\varepsilon}$, which will be more useful in the form $(Aw)(z) \leq \ell^{-1}(\sqrt{\varepsilon})$. Furthermore, since $U_- \subseteq \mathcal{D}^c$,

$$\tau^2 \leq \mu(U_-) = \mu(U_- \cap V) + \mu(U_- \cap V^c) \leq \mu(U_- \cap V) + \mu(\mathcal{D}^c \cap V^c) = \mu(U_- \cap V) + \mu(S_r),$$

which rearranges to give $\mu(U_- \cap V) \geq \tau^2 - \mu(S_r) \geq \tau^2 - \sqrt{\varepsilon} \geq \tau^2/2$. Note by Eq. (16) that

$$\mu(U \cap V_-) \geq \tau\mu(U_- \cap V) \geq \frac{\tau^3}{2},$$

and $z \in V_-$ has $(Aw)(z) \geq -\ell^{-1}(\sqrt{\varepsilon})$, and more importantly $\ell((Aw)(z)) \geq \ell(-\ell^{-1}(\sqrt{\varepsilon})) > 6\mathcal{R}(0)/\tau^3$. Consequently, since $\mathcal{E}(w) \leq \mathcal{R}(0)$ and $\mathcal{R}(0) > 0$,

$$\begin{aligned}
\mathcal{R}(w) &\geq \int_{U \cap V_-} \ell(Aw) d\mu(x,y) \\
&\geq \mu(U \cap V_-)\ell(-\ell^{-1}(\sqrt{\varepsilon})) \\
&\geq 3\mathcal{R}(0) \\
&\geq \mathcal{R}(0) + \mathcal{E}(w) + \inf_{v \in L_1(\mathcal{H})} \mathcal{R}(v) \\
&= \mathcal{R}(0) + \mathcal{R}(w) \\
&> \mathcal{R}(w),
\end{aligned}$$

a contradiction. ∎

Proof (of Lemma 3.6) First consider S_+; by $\ell \geq 0$ and convexity, $\ell(c_1) \geq \ell(0) + c_1\ell'(0) \geq c_1\ell'(0)$, and

$$\mathcal{R}(w) \geq \int_{S_+} \ell(Aw) d\mu \geq \ell(c_1)\mu(S_+) \geq c_1\ell'(0)\mu(S_+),$$

which rearranges to give $\mu(S_+) \leq \mathcal{R}(w)/(c_1\ell'(0))$.

To control S_- we take advantage of S_+: the region $\mathcal{D}$ is a set of points where it is impossible for S_- to be large without S_+ being large as well, and $\bar{q}$ is a witness to this fact. To start, note by $A^\top \bar{q} = 0$ and $\bar{q} = 0$ on $\mathcal{D}^c$ that

$$0 = \langle Aw, q\rangle = \int_{\mathcal{D}} (Aw)\bar{q}d\mu = \int_{Aw>0} (Aw)\bar{q}d\mu + \int_{Aw<0} (Aw)\bar{q}d\mu,$$

which rearranges to yield

$$\int_{Aw>0} (Aw)\bar{q}d\mu = -\int_{Aw<0} (Aw)\bar{q}d\mu.$$

Combining this with Hölder's inequality for Orlicz spaces (see Proposition B.1),

$$\begin{aligned} 2\|\bar{q}\|_{\beta^*} \left\|Aw\,\mathbf{1}[Aw>0]\right\|_\beta &\geq |\langle \bar{q}, Aw\,\mathbf{1}[Aw>0]\rangle| \\ &= |\langle \bar{q}, Aw\,\mathbf{1}[Aw<0]\rangle| \\ &\geq c_1c_2\mu(S_-). \end{aligned}$$

Now using the definition of $c_{\ell,\mu}$ and rearranging,

$$\mathcal{R}(w) \geq \int_{Aw>0} \ell(Aw)d\mu \geq \frac{c_1c_2\mu(S_-)}{2c_{\ell,\mu}\|\bar{q}\|_{\beta^*}},$$

which gives the desired bound on $\mu(S_-)$.

In order to control $|\bar{\eta} - \eta_w|$ on U, suppose without loss of generality that μ-a.e. $(x,y) \in \mathcal{D}$ satisfies $(x,-y) \in \mathcal{D}$ (see Lemma I.2), and define a scalar $L := \lim_{r\to\infty} \ell(r)$, a set $\mathcal{D}' := \{z \in \mathcal{D} : z < L\}$, and a function

$$\bar{f}(z) := \begin{cases} (\ell^*)'(\bar{q}(z)) & \text{when } z \in \mathcal{D}', \\ 0 & \text{otherwise,} \end{cases}$$

Note that $\bar{f}$ is well-defined (and measurable) by construction, since strict convexity of ℓ implies differentiability of ℓ^* along the interior of $\mathrm{dom}(\ell^*)$ (Hiriart-Urruty and Lemaréchal, 2001, Theorem E.4.1.1), which coincides with the set $\mathcal{D}'$ (because the domain of $(\ell^*)'$ is the image of ℓ' by first-order optimality for conjugates). By Taylor's theorem, for every $z \in \mathcal{D}'$ there exists $q_z \in [(Aw)_z, \bar{f}(z)]$ with

$$\begin{aligned} \ell\big((Aw)(z)\big) &= \ell(\bar{f}(z)) + \ell'(\bar{f}(z))\big((Aw)(z) - \bar{f}(z)\big) + \frac{1}{2}\big((Aw)(z) - \bar{f}(z)\big)^2\ell''(q_z) \\ &= -\ell^*(\bar{q}(z)) + \bar{q}(z)\big((Aw)(z)\big) + \frac{1}{2}\big((Aw)(z) - \bar{f}(z)\big)^2\ell''(q_z) \\ &\geq -\ell^*(\bar{q}(z)) + \bar{q}(z)\big((Aw)(z)\big) + \frac{\tau}{2}\big((Aw)(z) - \bar{f}(z)\big)^2\mathbf{1}[z \in U], \end{aligned}$$

where the second line made use of $\bar{q}(z) = \ell'(\bar{f}(z))$ and Fenchel's inequality. All terms in this final bound are integrable over $\mathcal{D}'$, and moreover either $\mathcal{D} = \mathcal{D}'$, or $L < \infty$ and $\mu(\mathcal{D} \setminus \mathcal{D}') = 0$ by

Lemma G.2, thus applying $\int_{\mathcal{D}}$ to both sides gives

$$\begin{aligned}\int_{\mathcal{D}} \ell(Aw)d\mu &\geq -\int_{\mathcal{D}} \ell^*(\bar{q})d\mu + \langle \bar{q}, Aw\rangle + \frac{\tau}{2}\int_U (Aw-\bar{f})^2 d\mu \\ &= \inf_{v\in L_1(\mathcal{H})} \int_{\mathcal{D}} \ell(Av)d\mu + \frac{\tau}{2}\int_U (Aw-\bar{f})^2 d\mu,\end{aligned}$$

which made use of $A^\top \bar{q} = 0$ and the fact that $\bar{q}$ also maximizes the dual problem restricted to $\mu_{\mathcal{D}}$ (by Lemma 3.2). Rearranging the preceding Taylor expansion gives

$$\int_U (Aw-\bar{f})^2 d\mu \leq \frac{2\mathcal{E}(w;\mu_{\mathcal{D}})}{\tau}. \tag{17}$$

The next step is to convert between $\bar{f}$ and $\bar{\eta}$. To this end, recall from the construction of $\bar{f}$ and subsequent discussion that $\bar{f}(z) = (\ell^*)'(\bar{q}(z))$ for μ-a.e. $z \in \mathcal{D}$ (and μ-a.e. $(x,y) \in \mathcal{D}$ has $(x,-y) \in \mathcal{D}$), thus Theorem 2.1.iv grants

$$\bar{f}(x,y) = (\ell^*)'(\bar{q}(x,y)) = -(\ell^*)'(\bar{q}(x,-y)) = -\bar{f}(x,-y) \qquad \text{for } \mu\text{-a.e. } (x,y) \in \mathcal{D}.$$

In particular, this grants $\phi(-\bar{f}(z)) = \bar{\eta}(z)$ for μ-a.e. $z \in \mathcal{D}$, which combined with Eq. (17) and the notation L_ϕ for the Lipschitz constant of ϕ means

$$\begin{aligned}\int_U |\eta - \eta_w|\, d\mu &= \int_U \left|\phi(-\bar{f}(z)) - \phi\big(-(Aw)(z)\big)\right| d\mu(z) \\ &\leq L_\phi \int_U \left|\bar{f} - Aw\right| d\mu \\ &\leq L_\phi \sqrt{\int_U \left|\bar{f} - Aw\right|^2 d\mu} \\ &\leq L_\phi \sqrt{\frac{2\mathcal{E}(w;\mu_{\mathcal{D}})}{\tau}},\end{aligned}$$

where the penultimate step used Jensen's inequality. ∎

Proof (of Theorem 1.1) First note that the bound for a single $w \in L_1(\mathcal{H})$ immediately implies the convergence result, thus it suffices to prove the bound.

To this end, let $w \in L_1(\mathcal{H})$ be given, set $\varepsilon := \mathcal{E}(w)$, and before defining f_1 (which will not depend on w), define two helper functions:

$$\begin{aligned}\tau(r) &= \inf_{|z|\leq r} \ell''(z)\ , \\ g_\varepsilon &= \begin{cases} \min\{r \geq 0 : \tau(r) \leq 2\sqrt{\varepsilon}\} & \text{if } 2\sqrt{\varepsilon} \leq \tau(1), \\ 1 & \text{if } 2\sqrt{\varepsilon} > \tau(1). \end{cases}\end{aligned}$$

The key properties are that $\tau(r) > 0$, it is continuous, non-increasing, and $\lim_{r\to\infty}\tau(r) = 0$ because $\liminf_{r\to-\infty}\ell''(r) = 0$ (by Lemma E.1). On the other hand, the definition of g_ε implies that

$$\tau(g_\varepsilon) = \min\left\{2\sqrt{\varepsilon},\, \tau(1)\right\}\ ,$$

which means that $g_\varepsilon \to \infty$ as $\varepsilon \downarrow 0$.

Next, f_1 will be constructed by splitting $\|\bar\eta - \eta_w\|_1$ along $\mathcal{D}$ and $\mathcal{D}^c$, and subsequently using Lemma 3.6 and Lemma 3.5 to control each term. When applying Lemma 3.5, the bound may be simplified by using $r := \sqrt{\varepsilon}$ and $\mu(\mathcal{D}^c \setminus S_r) \le 1$. When applying Lemma 3.6 (and using Corollary 3.3 to give $\mathcal{E}(w;\mu_{\mathcal{D}}) \le \varepsilon$), the bound may be simplified by setting $c_1 := g_\varepsilon$, $c_2 := \max\{c_1^{-1/2}, \ell'(-g_\varepsilon)\}$, and $c_3 := \ell'(g_\varepsilon)$. With these definitions, it follows that the τ of Lemma 3.6, which equals $\min\{\inf_{|z|\le c_1} \ell''(z), \inf_{z\in[c_2,c_3]} \ell''((\ell^*)'(z))\}$, coincides with $\tau(g_\varepsilon)$. If $c_3 < c_2$, set $f_1(\varepsilon) = 1$; otherwise, Lemma 3.6 may be applied, and together with the terms from Lemma 3.5 it follows that

$$\begin{aligned}
\int |\bar\eta - \eta_w| d\mu &= \int_{\mathcal{D}^c} |\bar\eta - \eta_w| d\mu + \int_{\mathcal{D}} |\bar\eta - \eta_w| d\mu \\
&\le \sqrt{\varepsilon} + \sqrt{\varepsilon}\max\left\{\frac{1}{\ell(0)}, \frac{c_\ell}{\ell'(0)}\right\} \\
&\quad + \left(\varepsilon + \inf_{v\in L_1(\mathcal{H})} \int \ell(Av) d\mu\right)\left(\frac{1}{g_\varepsilon \ell'(0)} + \frac{2c_{\ell,\mu}\|\bar q\|_{\beta^*}}{\sqrt{g_\varepsilon}}\right) \\
&\quad + \underbrace{L_\phi \sqrt{\frac{2\varepsilon}{\tau(g_\varepsilon)}}}_{\star} \\
&\quad + \underbrace{\mu(\{z \in \mathcal{Z} : \bar q(z) \in (0, \max\{\ell'(-g_\varepsilon), g_\varepsilon^{-1/2}\}) \vee \bar q(z) > \ell'(g_\varepsilon)\})}_{\triangle} \\
&=: f_1(\varepsilon).
\end{aligned}$$

By construction, f_1 is well-defined, does not depend on w, and satisfies the desired inequality; it remains to be shown that $f_1(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$. It suffices to consider $\triangle$ and $\star$, since all other terms contain ε in a numerator, or g_ε in a denominator (where, as shown before, $g_\varepsilon \to \infty$ as $\varepsilon \downarrow 0$), without any worry of cancellations mitigating these effects.

To handle $\triangle$, first expand the terms as

$$\triangle \le \underbrace{\mu(\{z \in \mathcal{Z} : \bar q(z) \in (0, \max\{\ell'(-g_\varepsilon), g_\varepsilon^{-1/2}\})\})}_{\square} + \underbrace{\mu(\{z \in \mathcal{Z} : \bar q(z) > \ell'(g_\varepsilon)\})}_{\lozenge}.$$

$\square \to 0$ as $\varepsilon \downarrow 0$, since $g_\varepsilon^{-1/2} \to 0$ as $\varepsilon \downarrow 0$ and since $\ell'(-r) \to 0$ as $r \to \infty$ by Lemma E.1. Lastly, to show $\lozenge \to 0$, there are two cases. First, if ℓ' grows unboundedly, then $\ell'(g_\varepsilon)$ will cover all values as $\varepsilon \downarrow 0$. On the other hand, if $L := \lim_{r\to\infty} \ell'(r) < \infty$, then $\mu(\{z \in \mathcal{Z} : \bar q \ge L\}) = 0$ as provided by Lemma G.2 means once again that $\ell'(g_\varepsilon)$ will cover all values (μ-a.e.) as $\varepsilon \downarrow 0$.

Lastly, to handle $\star$, we use $\tau(g_\varepsilon) = \min\{2\sqrt{\varepsilon}, \tau(1)\}$ to obtain that

$$L_\phi \sqrt{2\varepsilon/\tau(g_\varepsilon)} = L_\phi \max\{\sqrt{2\varepsilon/\tau(1)}, \varepsilon^{1/4}\} \ ,$$

which goes to zero as $\varepsilon \to 0$. ■

Appendix J. Proofs from Section 3.2

As in the main text, this appendix first develops the quantity $\mathrm{Bal}(\mu)$, and then uses it to develop the deviation bounds.

J.1. Basic properties of Bal(μ)

To start, note the range of values for $\mathrm{Bal}(\mu)$. When $\mathrm{Ker}(\mu)^\perp \neq \{0\}$, then $\mathrm{Bal}(\mu) \in [0, \mu(\mathcal{Z})]$, but the case $\mathrm{Ker}(\mu)^\perp = \{0\}$ means, via usual conventions on infima, that $\mathrm{Bal}(\mu) = \infty$. This represents a certain degeneracy in the learning problem; indeed, it is a scenario where there is nothing to learn, since equivalently $\mathrm{Ker}(\mu) = \mathbb{R}^d$, and thus every element of $\mathbb{R}^d$ has no impact on the problem.

With this in mind, the first lemma relates boundedness and risk.

Lemma J.1 *Let finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, and $\bar{s} \in \partial\ell(0)$ be given. Then every $w \in \mathrm{Ker}(\mu)^\perp$ satisfies*

$$\|w\|_1 \leq \frac{\mathcal{R}(w)}{\bar{s}\mathrm{Bal}(\mu)}.$$

Proof By definition of $\mathrm{Bal}(\mu)$, since $\ell \geq 0$,

$$\mathcal{R}(w) \geq \int_{Aw>0} \ell(Aw)d\mu \geq \int_{Aw>0} (\ell(0) + \bar{s}(Aw))d\mu \geq \|w\|_1 \bar{s}\mathrm{Bal}(\mu),$$

which rearranges to give the result. (As a sanity check, the case $\mathrm{Bal}(\mu) = \infty$ means $\mathrm{Ker}(\mu)^\perp = \{0\}$, whereby $\|w\|_1 = 0$ automatically.) ■

Next, note that the infimand within the definition of $\mathrm{Bal}(\mu)$ is Lipschitz continuous.

Lemma J.2 *Let finite measure μ and hypotheses $\mathcal{H}$, $|\mathcal{H}| = d$, be given, and define the function $f(w) := \int_{Aw>0}(Aw)d\mu$. Then, for every $w, w' \in \mathbb{R}^d$,*

$$|f(w) - f(w')| \leq \|w - w'\|_1 \mu(\mathcal{Z}).$$

Proof Let w, w' be given, and define $N(v) := \{z \in \mathcal{Z} : (Av)(z) > 0\}$. Since $|h(z)| \leq 1$ for every $h \in \mathcal{H}$ (whereby $|(Av)(z)| \leq \|v\|_1$ for every z and v),

$$\begin{aligned}
|f(w) - f(w')| &= \left| \int_{N(w)\cap N(w')} A(w - w')d\mu + \int_{N(w)\setminus N(w')} (Aw)d\mu - \int_{N(w')\setminus N(w)} (Aw')d\mu \right| \\
&\leq \|w - w'\|_1 \mu(N(w) \cap N(w')) + \left| \int_{N(w)\setminus N(w')} (Aw)d\mu \right| + \left| \int_{N(w')\setminus N(w)} (Aw')d\mu \right|.
\end{aligned}$$

Since the second and third terms are symmetric, it suffices to consider the second. To this end, note that

$$z \in N(w) \setminus N(w') \quad \Longrightarrow \quad (Aw)(z) > 0$$

and

$$z \in N(w) \setminus N(w') \quad \Longrightarrow \quad (Aw)(z) = (Aw')(z) + (A(w - w'))(z) \leq 0 + \|w - w'\|_1,$$

which combine to give

$$z \in N(w) \setminus N(w') \quad \Longrightarrow \quad |(Aw)(z)| \leq \|w - w'\|_1,$$

and thus

$$\left| \int_{N(w)\setminus N(w')} (Aw) d\mu \right| \leq \int_{N(w)\setminus N(w')} |Aw| \, d\mu \leq \|w - w'\|_1 \mu(N(w) \setminus N(w')).$$

The result follows. ■

It will now be shown that $\mathrm{Bal}(\mu_D) > 0$ whenever $\mu_D > 0$. To prove this, the preceding lemma showed that the infimand in the definition of $\mathrm{Bal}(\mu)$ is continuous; on the other hand, since $|\mathcal{H}| < \infty$, the domain of the infimum is compact, which together with the aforementioned continuity gives attainment at a necessarily positive point.

Lemma J.3 *Let finite measure μ, hypotheses $\mathcal{H}$, loss $\ell \in \mathbb{L}$, and dual variable $q \in L_{\beta^*}(\mu)$ with $q \geq 0$ μ-a.e. and $A^\top q = 0$ be given, and set $D := \{z \in \mathcal{Z} : q(z) > 0\}$. If $\mu(D) > 0$ and $|\mathcal{H}| < \infty$, then $\mathrm{Bal}(\mu_D) > 0$.*

Proof If $\mathrm{Ker}(\mu)^\perp = \{0\}$, then $\mathrm{Bal}(\mu) = \infty > 0$ immediately, thus suppose $\mathrm{Ker}(\mu)^\perp$ is a nontrivial subspace, meaning in particular that there exists $w \in \mathrm{Ker}(\mu)^\perp$ with $\|w\|_1 = 1$. By Lemma J.2, the map $w \mapsto \int_{Aw>0}(Aw)d\mu_D$ is continuous; since moreover the (nonempty) set $C = \{w \in \mathrm{Ker}(\mu)^\perp : \|w\|_1 = 1\}$ is compact when $|\mathcal{H}| < \infty$, it follows that the minimization in the definition of $\mathrm{Bal}(\mu_D)$ is attained at some point in C. The remainder of the proof establishes that the integral is indeed positive everywhere on C.

Consider any $w \in C$. Since $C \cap \mathrm{Ker}(\mu_D) = \emptyset$, it must hold that $\mu_D(\{z \in \mathcal{Z} : (Aw)(z) \neq 0\}) > 0$ (else $w \in \mathrm{Ker}(\mu_D)$), and thus at least one of the two expressions $\int_{Aw>0}(Aw)d\mu_D$ and $\int_{Aw<0}(Aw)d\mu_D$ must be nonzero. If the first is nonzero, it is positive, and the proof is done, thus suppose that only the second expression is nonzero, which necessarily means it is negative. Since $A^\top q = 0$, then $\langle Aw, q\rangle = 0$, which can be split into negative and positive parts to yield

$$\int_{Aw>0} (Aw) q \, d\mu_D = -\int_{Aw<0} (Aw) q \, d\mu_D > 0$$

as desired. ■

Lemma J.3 was stated for general $q \in L_{\beta^*}(\mu)$ due to its use in future lemmas; however, by instantiating it for a dual optimum $\bar{q}$, it follows that $\mathrm{Bal}(\mu_{\mathcal{D}}) > 0$ whenever $\mu(\mathcal{D}) > 0$.

Proposition J.4 *Let μ be a finite measure, $\mathcal{H}$ be a hypothesis set with $|\mathcal{H}| < \infty$, and $\ell \in \mathbb{L}$ be a loss with corresponding difficult set $\mathcal{D}$. Then $\mathrm{Bal}(\mu_{\mathcal{D}}) > 0$ whenever $\mu(\mathcal{D}) > 0$.*

Proof The result follows by applying Lemma J.3 to $\bar{q}$ and $\mathcal{D}$, noting that they satisfy the desired properties by Theorem 2.1 and the definition of $\mathcal{D}$. ■

The next two properties will establish the interplay between Bal, $\mathcal{D}$, and also primal-dual optimal pairs $(\bar{w}, \bar{q})$.

Lemma J.5 *Let finite measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| < \infty$, loss $\ell \in \mathbb{L}$, and $\bar{s} \in \partial\ell(0)$ be given. If* $\mathrm{Bal}(\mu) > 0$*, then there exists a primal-dual optimal pair $(\bar{w}, \bar{q})$ to Eq. (7) which satisfies $\bar{w} \in \mathrm{Ker}(\mu)^\perp$, and $\|\bar{w}\|_1 \leq \ell(0)\mu(\mathcal{Z})/(\bar{s}\mathrm{Bal}(\mu))$, and $\bar{q}(z) \in \partial\ell((Aw)(z))$ for μ-a.e. $z \in \mathcal{Z}$.*

Proof From the definition of $\mathrm{Ker}(\mu)$, it suffices to optimize the primal over $\mathrm{Ker}(\mu)^\perp$, and by Lemma E.1, the primal optimization can be further restricted to the compact convex set

$$\left\{ w \in \mathrm{Ker}(\mu)^\perp : \ \|w\|_1 \leq \ell(0)\mu(\mathcal{Z})/(\bar{s}\mathrm{Bal}(\mu)) \right\} \ ,$$

where a minimum $\bar{w}$ is attained (by continuity of convex functions on $\mathbb{R}^d$). The relationship with $\bar{q}$ follows by Theorem 2.1. ■

The remainder of this subsection will build towards the construction of the canonical difficult set $\mathcal{D}_\star$: the difficult sets $\mathcal{D}$ provided by losses in $\mathbb{L}^{2+}$ are "maximal" in the measure-theoretic sense. To this end, the following lemma is essential.

Lemma J.6 *Let finite measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| < \infty$, loss $\ell \in \mathbb{L}^{2+}$, and a corresponding difficult set $\mathcal{D}$ be given. For any set S with* $\mathrm{Bal}(\mu_S) > 0$*, then $\mu(S \setminus \mathcal{D}) = 0$.*

Proof Suppose contradictorily that $\mu(S \setminus \mathcal{D}) > 0$, which entails $\mu(S) > 0$, and let $\bar{q}$ denote the dual optimum associated with $\mathcal{D}$.

Applying Lemma J.5 to loss ℓ and measure μ_S, it follows from $\mathrm{Bal}(\mu_S) > 0$ that there exists a primal optimum $\bar{w}_S$ and corresponding dual optimum $\bar{q}_S$ with $\bar{q}_S \in \partial\ell(A\bar{w}_S)$ μ_S-a.e., and consequently $\bar{q}_S > 0$ μ-a.e. since $\ell \in \mathbb{L}^{2+}$.

Define $\hat{q}(z) := \bar{q}(z) + \bar{q}_S(z)\mathbf{1}[z \in S]$, whereby, for any $w \in \mathbb{R}^d$,

$$\langle Aw, \hat{q} \rangle = \langle Aw, \bar{q} \rangle + \langle Aw, \bar{q}_S \mathbf{1}[z \in S] \rangle = \int (Aw)(\bar{q}) d\mu + \int_S (Aw)\bar{q}_S d\mu = 0 + 0.$$

Additionally, $\hat{q} \geq 0$ μ-a.e. with $\hat{q} > 0$ μ-a.e. along $D := \mathcal{D} \cup S$, thus Lemma J.3 and Lemma J.5 may be applied to obtain a dual optimum $\bar{q}_D$ which is positive μ-a.e. along D. Of course, $\bar{q}$ was feasible for the problem restricted to D, and by strict convexity of $\int \ell^*(q)d\mu$ (see Proposition A.3), it follows that $\int \ell^*(\bar{q}_D)d\mu_D < \int \ell^*(\bar{q})d\mu_D$. But this is a contradiction, since $z \mapsto \bar{q}_D(z)\mathbf{1}[z \in D]$ is feasible for the full problem without changing its objective value, and meanwhile $\bar{q}$ was optimal for the full problem. ■

Proof (of Proposition 3.9) We will show that if $\ell_1 \in \mathbb{L}$ and $\ell_2 \in \mathbb{L}^{2+}$ with corresponding difficult sets $\mathcal{D}_1$ and $\mathcal{D}_2$, then $\mu(\mathcal{D}_1 \setminus \mathcal{D}_2) = 0$, which will yield the proof. For (i), it suffices to instantiate the claim with $\ell_1 = \ell$ and $\ell_2 = \exp$, and (ii) follows by instantiating the claim once with $\ell_1 = \ell$ and $\ell_2 = \exp$, and a second time with $\ell_1 = \exp$ and $\ell_2 = \ell$.

The proof of the general claim is as follows. If $\mu(\mathcal{D}_1) = 0$, then $\mu(\mathcal{D}_1 \setminus \mathcal{D}_2) = 0$ automatically, thus suppose $\mu(\mathcal{D}_1) > 0$. In this case, Lemma J.3 grants $\mathrm{Bal}(\mu_{\mathcal{D}_1}) > 0$, and thus applying Lemma J.6 with loss ℓ_2 and $S := \mathcal{D}_1$ gives $\mu(\mathcal{D}_1 \setminus \mathcal{D}_2) = 0$. ■

J.2. Splitting $\widehat{\mathcal{R}}_n$ along $\mathcal{D}$ and $\mathcal{D}^c$

As granted by the development of $\mathrm{Bal}(\mu)$, recall from the main text that there exists a canonical difficult set $\mathcal{D}_\star$, which by Proposition 3.9 is not tied to any specific loss. The goal of this section is to show, as stated in Lemma J.8, that $\widehat{\mathcal{R}}$ can be split along $\mathcal{D}_\star$, just like $\mathcal{R}$ (cf. Corollary 3.3), despite $\mathcal{D}_\star$ being constructed over μ rather than $\widehat{\mu}$.

As the first step, we establish the existence of arbitrarily good predictors over $\mathcal{D}_\star^c$.

Lemma J.7 *Let finite measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| = d$, and canonical difficult set $\mathcal{D}_\star$ be given. Then for every $\varepsilon > 0$, there exists $v \in \mathbb{R}^d$ such that $(Av)(z) = 0$ for μ-a.e. $z \in \mathcal{D}_\star$, and $\mu(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\}) \leq \varepsilon$.*

Proof Throughout this proof, set $\ell := \exp \in \mathbb{L}_{\mathrm{b}}^{2+}$, whereby $\mathcal{D} = \mathcal{D}_\star$ by definition of $\mathcal{D}_\star$, and let $\varepsilon > 0$ be given.

There are now two cases to consider; first consider the simpler case $\mu(\mathcal{D}_\star) = 0$. Choose any $v \in L_1(\mu)$ with $\mathcal{E}(v) \leq \varepsilon\ell(-1)$, and first note that $(Av)(z) = 0$ for μ-a.e. $z \in \mathcal{D}_\star$ without any effort since $\mu(\mathcal{D}_\star) = 0$. On the other hand, by Lemma 3.5 (with $r := \ell(-1)$),

$$\mu(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\}) = \mu(\{z \in \mathcal{D}_\star^c : \ell((Av)(z)) \geq \ell(-1)\}) \leq \frac{\mathcal{E}(w)}{\ell(-1)} \leq \varepsilon,$$

which completes the proof under the assumption $\mu(\mathcal{D}_\star) = 0$.

Now consider the case $\mu(\mathcal{D}_\star) > 0$, whereby Proposition J.4 grants $\mathrm{Bal}(\mu_{\mathcal{D}_\star}) > 0$. Let $\bar{s} \in \partial\ell(0)$ be arbitrary and set

$$B := \frac{1 + \inf_{v \in \mathbb{R}^d} \mathcal{R}(v; \mu_{\mathcal{D}_\star})}{\bar{s}\mathrm{Bal}(\mu_{\mathcal{D}_\star})},$$

whereby Lemma J.1 grants that every $w \in \mathrm{Ker}(\mu_{\mathcal{D}_\star})^\perp$ with $\mathcal{E}(w, \mu_{\mathcal{D}_\star}) \leq 1$ satisfies

$$\|w\|_1 \leq \frac{\mathcal{R}(w; \mu_{\mathcal{D}_\star})}{\bar{s}\mathrm{Bal}(\mu_{\mathcal{D}_\star})} \leq B.$$

Now let $\varepsilon > 0$ be given, and choose $\varepsilon_0 \in (0, \min\{\varepsilon^2, 1\}]$ such that $\ell^{-1}(\sqrt{\varepsilon_0}) \leq -1 - B$. Let $u \in \mathbb{R}^d$ be given with $\mathcal{E}(u) \leq \varepsilon_0$, whereby Corollary 3.3 grants that $\max\{\mathcal{E}(u; \mu_{\mathcal{D}_\star}), \mathcal{R}(u; \mu_{\mathcal{D}_\star^c})\} \leq \varepsilon_0$ as well. By Lemma 3.5 with $r := \sqrt{\varepsilon_0}$ and the above definitions,

$$\begin{aligned}
\varepsilon \geq \sqrt{\varepsilon_0} &\geq \mu(\{z \in \mathcal{D}_\star^c : \ell((Au)(z)) \geq \sqrt{\varepsilon_0}\}) \\
&= \mu(\{z \in \mathcal{D}_\star^c : (Au)(z) \geq \ell^{-1}(\sqrt{\varepsilon_0})\}) \\
&\geq \mu(\{z \in \mathcal{D}_\star^c : (Au)(z) \geq -1 - B\}).
\end{aligned}$$

Now write u as the direct sum $u = v \oplus u_\perp$, where $v \in \mathrm{Ker}(\mu_{\mathcal{D}_\star})$ and $u_\perp \in \mathrm{Ker}(\mu_{\mathcal{D}_\star})^\perp$. By the earlier derivation, $\|u_\perp\|_1 \leq B$, and thus, for any $z \in \mathcal{Z}$, we have $|(Au_\perp)(z)| \leq B$, and

$$(Av)(z) \geq -1 \quad \Longrightarrow \quad (Av)(z) \geq -1 - B - (Au_\perp)(z) \quad \Longleftrightarrow \quad (Au)(z) \geq -1 - B.$$

This combines with the earlier derivation to yield

$$\varepsilon \geq \mu(\{z \in \mathcal{D}_\star^c : (Au)(z) \geq -1 - B\}) \geq \mu(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\})$$

as desired. ■

Thanks to the preceding lemma, splitting $\widehat{\mathcal{R}}_n$ into $\mathcal{D}_\star$ and $\mathcal{D}_\star^c$, is straightforward (and similar to the proof of Corollary 3.3).

Lemma J.8 *Let probability measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| = d$, and canonical difficult set $\mathcal{D}_\star$ be given. Then, for any $\delta > 0$, with probability $1 - \delta$ over a random draw of size n from μ, every loss $\ell \in \mathbb{L}^{2+}$ satisfies*

$$\inf_{w \in \mathbb{R}^d} \mathcal{R}(w; \widehat{\mu}) = \inf_{w \in \mathbb{R}^d} \mathcal{R}(w; \widehat{\mu}_{\mathcal{D}_\star})$$

and for every $w \in \mathbb{R}^d$

$$\mathcal{E}(w; \widehat{\mu}_{\mathcal{D}_\star}) \leq \mathcal{E}(w; \widehat{\mu}) \ , \quad \mathcal{R}(w; \widehat{\mu}_{\mathcal{D}_\star^c}) \leq \mathcal{E}(w; \widehat{\mu}) \ .$$

Proof Let $\mathcal{S}$ to denote the sample, where $\mathcal{S}_c := \mathcal{S} \cap \mathcal{D}_\star^c$ with size $n_c := |\mathcal{S}_c|$ denotes the portion falling within $\mathcal{D}_\star^c$, and $\mathcal{S}_{\mathcal{D}} := \mathcal{S} \cap \mathcal{D}_\star$ the portion falling within $\mathcal{D}_\star$. If $n_c = 0$, then all claims follow immediately (indeed, this implies $\widehat{\mu} = \widehat{\mu}_{\mathcal{D}_\star}$ and $\mathcal{R}(w; \widehat{\mu}_{\mathcal{D}_\star^c}) = 0$), thus suppose $n_c > 0$.

By Lemma J.7 with $\varepsilon := \mu(\mathcal{D}_\star^c)\varepsilon_0$ and $\varepsilon_0 := \min\{1/2, -\ln(1-\delta)/(2n_c)\}$, there exists $v \in \mathbb{R}^d$ satisfying $(Av)(z) = 0$ for μ-a.e. $z \in \mathcal{D}_\star$, and

$$\mu(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\}) \leq \varepsilon = \mu(\mathcal{D}_\star^c)\varepsilon_0,$$

or equivalently

$$\mu_{|\mathcal{D}_\star^c}(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\}) \leq \varepsilon_0 \ .$$

Consequently, with probability at least $1 - \delta$ over the draw of $\mathcal{S}$, conditional on n_c, we obtain $(Av)(z_i) = 0$ for every $z_i \in \mathcal{S}_{\mathcal{D}}$ and $(Av)(z_i) \leq -1$ for every $z_i \in \mathcal{S}_c$, the latter statement since

$$\begin{aligned}
\Pr[\forall i \in \mathcal{S}_c, \, (Av)(z_i) \leq -1] &= \mu_{|\mathcal{D}_\star^c}(\{z \in \mathcal{D}_\star^c : (Av)(z) \geq -1\})^{n_c} \\
&\geq (1 - \varepsilon_0)^{n_c} \\
&\geq (1 - (2\varepsilon_0) + (2\varepsilon_0)^2/2)^{n_c} \\
&\geq \exp(-2n_c\varepsilon_0) \\
&\geq 1 - \delta.
\end{aligned}$$

Since $\lim_{z \to -\infty} \ell(z) = 0$, every $w \in \mathbb{R}^d$ and $z_i \in \mathcal{S}_c$ satisfies $\inf_{r \geq 0} \ell((A(w + rv))(z_i)) = 0$, thus

$$\begin{aligned}
\inf_{w \in \mathbb{R}^d} \mathcal{R}(w; \widehat{\mu}) &= \inf_{\substack{w \in \mathbb{R}^d \\ r \geq 0}} \int \ell(A(w + rv)) d\widehat{\mu} \\
&= \inf_{\substack{w \in \mathbb{R}^d \\ r \geq 0}} \left(\int_{\mathcal{D}_\star} \ell(A(w + rv)) d\widehat{\mu} + \int_{\mathcal{D}_\star^c} \ell(A(w + rv)) d\widehat{\mu} \right) \\
&= \inf_{w \in \mathbb{R}^d} \left(\int_{\mathcal{D}_\star} \ell(Aw) d\widehat{\mu} + \inf_{r \geq 0} \int_{\mathcal{D}_\star^c} \ell(A(w + rv)) d\widehat{\mu} \right) \\
&= \inf_{w \in \mathbb{R}^d} \mathcal{R}(w; \widehat{\mu}_{\mathcal{D}_\star}).
\end{aligned}$$

For the last part, proceeding similarly to the proof of Lemma 3.2, the above derivation and $\ell \geq 0$ grant

$$\mathcal{E}(w;\widehat{\mu}_{\mathcal{D}_\star}) = \int_{\mathcal{D}_\star} \ell(Aw)d\widehat{\mu} - \inf_{v\in\mathbb{R}^d} \int_{\mathcal{D}_\star} \ell(Av)d\widehat{\mu} \leq \int \ell(Aw)d\widehat{\mu} - \inf_{v\in\mathbb{R}^d} \int \ell(Av)d\widehat{\mu} = \mathcal{E}(w;\widehat{\mu})$$

directly, and

$$\begin{aligned}
\mathcal{R}(w;\widehat{\mu}_{\mathcal{D}_\star^c}) &= \int \ell(Aw)d\widehat{\mu} - \int_{\mathcal{D}_\star} \ell(Aw)d\widehat{\mu} \\
&\leq \int \ell(Aw)d\widehat{\mu} - \inf_{v\in\mathbb{R}^d} \int_{\mathcal{D}_\star} \ell(Av)d\widehat{\mu} \\
&= \int \ell(Aw)d\widehat{\mu} - \inf_{v\in\mathbb{R}^d} \int \ell(Av)d\widehat{\mu} = \mathcal{E}(w;\widehat{\mu}).
\end{aligned}$$

■

J.3. Controlling deviations over $\mathcal{D}_\star^c$

This section will establish the deviation bound over $\mathcal{D}_\star^c$, namely Lemma J.9. Superficially, this is merely an application of the VC theorem, however there are two issues under the surface.

First, note that this lemma does not attempt to control $\mathcal{E}(w;\mu_{\mathcal{D}_\star^c})$, which of course would allow a direct application of Lemma 3.5 and ostensibly an easy analysis over $\mathcal{D}_\star^c$ within the proof of Theorem 1.2. The reason is that there is evidence $\mathcal{E}(w;\mu_{\mathcal{D}_\star^c})$ cannot be controlled without placing strong restrictions on $w \in \mathbb{R}^d$ (Levy et al., 2014). On the other hand, the margin-like bound here is sufficient to aid in the proof of Theorem 1.2.

The second issue is that $\mathcal{D}_\star^c$ is an object constructed over μ rather than $\widehat{\mu}$, which is circumvented via Lemma J.8.

Lemma J.9 *Let probability measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| = d$, loss $\ell \in \mathbb{L}$, and canonical difficult set $\mathcal{D}_\star$ with $\mu(\mathcal{D}_\star^c) > 0$ be given. Then with probability at least $1-2\delta$ over an i.i.d. draw of size n from $\mu_{|\mathcal{D}_\star^c}$, every $w \in \mathbb{R}^d$ and $\varepsilon > 0$ with $\varepsilon \geq \sqrt{\hat{\mathcal{E}}_n(w)}$ satisfies*

$$\mu_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : \ell(Aw) \geq \varepsilon\}\right) \leq \frac{\sqrt{\hat{\mathcal{E}}_n(w)}}{\widehat{\mu}(\mathcal{D}_\star^c)} + \sqrt{\frac{32(1+d)\ln(1+n) + 4\ln(1/\delta)}{n}}.$$

Proof Since the set of linear threshold functions with weight vectors in $\mathbb{R}^d$ has VC dimension $1+d$, the nondecreasing property of ℓ combined with the VC theorem grants (Boucheron et al., 2005), with probability $1-\delta$,

$$\begin{aligned}
&\sup_{\substack{w\in\mathbb{R}^d \\ s>0}} \left|\mu_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : \ell(Aw) \geq s\}\right) - \widehat{\mu}_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : \ell(Aw) \geq s\}\right)\right| \\
&\qquad \leq \sup_{\substack{w\in\mathbb{R}^d \\ r\in\mathbb{R}}} \left|\mu_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : (Aw) \geq r\}\right) - \widehat{\mu}_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : (Aw) \geq r\}\right)\right| \\
&\qquad \leq \sqrt{\frac{16(1+d)\ln(1+n) + 2\ln(1/\delta)}{n}}.
\end{aligned}$$

Now let $w \in \mathbb{R}^d$ be arbitrary. Instantiating the above display with w and $s := \varepsilon > 0$, and then applying Lemma I.1 on measure $\widehat{\mu}_{|\mathcal{D}_\star^c}$ and set $\mathcal{D}_\star^c$ with scalar $r := \varepsilon > 0$, it follows that

$$\mu_{|\mathcal{D}_\star^c}\left(\{z \in \mathcal{D}_\star^c : \ell(Aw) \geq \varepsilon\}\right) \leq \frac{\mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star^c})}{\varepsilon} + \sqrt{\frac{16(1+d)\ln(1+n) + 2\ln(1/\delta)}{n}}.$$

To finish, Lemma J.8 grants $\mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star^c}) = \mathcal{R}(w; \widehat{\mu}_{\mathcal{D}_\star^c})/\widehat{\mu}(\mathcal{D}_\star^c) \leq \hat{\mathcal{E}}_n(w)/\widehat{\mu}(\mathcal{D}_\star^c)$ after discarding another δ failure probability, and the result follows by plugging in $\varepsilon \geq \sqrt{\hat{\mathcal{E}}_n(w)}$. ■

J.4. Controlling deviations over $\mathcal{D}_\star$: Proof of Lemma 3.10

In order to establish Lemma 3.10, two lemmas are in order: the first shows that $\mathrm{Bal}(\mu)$ is statistically stable, and the second develops a refined deviation bound for $\mathbb{L}_{\mathrm{b}}^{2+}$ over $\mathcal{D}_\star$.

Lemma J.10 *Let probability measure μ, hypotheses $\mathcal{H}$ with $|\mathcal{H}| = d$, and canonical difficult set $\mathcal{D}_\star$ with $\mu(\mathcal{D}_\star) > 0$ be given. Then with probability at least $1 - \delta$ over a draw of size n from $\mu_{|\mathcal{D}_\star}$, $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \subseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, and*

$$\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) \geq \mathrm{Bal}(\mu_{|\mathcal{D}_\star}) - 8\sqrt{\frac{\ln(2d) + \ln(4/\delta)}{n}}.$$

Moreover, if $n \geq 256(\ln(2d+\ln(4/\delta))/\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2$, then $\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) \geq \mathrm{Bal}(\mu_{|\mathcal{D}_\star})/2$ and $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) = \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$.

Proof Let $\mathcal{S} := (z_i)_{i=1}^n$ denote the random draw from $\mu_{|\mathcal{D}_\star}$.

First it will be shown that $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \subseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$ with probability 1. If $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) = \{0\}$, then the claim is immediate, thus suppose $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})$ is a nontrivial subspace of $\mathbb{R}^d$. Pick an orthonormal basis $(w_j)_{j=1}^k$ for $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})$. For each w_j, define $N_j := \{z \in \mathcal{Z} : (Aw_j)(z) \neq 0\}$, whereby $\mu_{|\mathcal{D}_\star}(N_j) = 0$ since $w_j \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})$. Since $\mu(\cup_{j=1}^k N_j) = 0$, then with probability 1 over the draw of sample $\mathcal{S}$, every $z_i \in \mathcal{S}$ and w_j satisfy $(Aw_j)(z_i) = 0$. Consequently, given an arbitrary $w \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})$, there exist scalars $(\alpha_j)_{j=1}^k$ such that $w = \sum_{j=1}^k \alpha_j w_j$, and thus, for every $z_i \in \mathcal{S}$, by linearity

$$(Aw)(z_i) = \sum_{j=1}^k \alpha_j (Aw_j)(z_i) = 0,$$

meaning $w \in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$ as well. Hence, $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \subseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$.

Throughout the remainder of this proof, discard the failure event for the above control on $\mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$: in particular, suppose $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \subseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, and equivalently $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp \supseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$.

In order to produce the lower bound on $\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star})$, first consider the case $\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) = \infty$. This means $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp = \{0\}$, which combined with $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp \supseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$ means $\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) = \infty$ as well, giving the desired bound.

Now consider the case $\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) < \infty$. First note that the map $z \mapsto \max\{(Aw)(z), 0\}$ is the composition of the 1-Lipschitz univariate map $\max\{\cdot, 0\}$ together with a linear function, so by Lemma C.2, it has Rademacher complexity $\|w\|_1\sqrt{2\ln(2d)/n}$ since $|-yh(x)| \leq 1$ for all (x, y).

Combining this with standard deviation bounds for Rademacher complexity (Lemma C.1), with probability $1-\delta$,

$$\sup_{\|w\|\leq 1}\left(\int \max\{Aw,0\}d\mu_{|\mathcal{D}_\star} - \int \max\{Aw,0\}d\widehat{\mu}_{|\mathcal{D}_\star}\right) \leq 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}}. \tag{18}$$

Combining Eq. (18) with $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp \supseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$,

$$\begin{aligned}
\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) &= \inf\left\{\int \max\{Aw,0\}d\widehat{\mu}_{|\mathcal{D}_\star} : \|w\|_1 = 1, w\in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp\right\} \\
&\geq \inf\left\{\int \max\{Aw,0\}d\mu_{|\mathcal{D}_\star} : \|w\|_1 = 1, w\in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp\right\} - 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}} \\
&\geq \inf\left\{\int \max\{Aw,0\}d\mu_{|\mathcal{D}_\star} : \|w\|_1 = 1, w\in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp\right\} - 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}} \\
&= \mathrm{Bal}(\mu_{|\mathcal{D}_\star}) - 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}}.
\end{aligned}$$

For the last statements, suppose the provided lower bound on n; this immediately grants the bound $\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) \geq \mathrm{Bal}(\mu_{|\mathcal{D}_\star})/2$ by the above derivation. To show $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) = \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, it suffices, by the above, to show $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \supseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$. If $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp = \{0\}$, then $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp \subseteq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$ immediately since the latter is a subspace, thus suppose $\mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$ is nontrivial. Let $w\in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$ with $\|w\|_1 > 0$ be arbitrary; by the definition of $\mathrm{Bal}(\mu_{|\mathcal{D}_\star})$, the lower bound on n, the deviation bound from Eq. (18), and since $\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) > 0$ because $\mu(\mathcal{D}_\star) > 0$ (see Proposition J.4),

$$\begin{aligned}
\int \max\{Aw,0\}d\widehat{\mu}_{|\mathcal{D}_\star} &= \|w\|_1\left(\int \max\{Aw/\|w\|_1,0\}d\widehat{\mu}_{|\mathcal{D}_\star}\right) \\
&\geq \|w\|_1\left(\int \max\{Aw/\|w\|_1,0\}d\mu_{|\mathcal{D}_\star} - 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}}\right) \\
&\geq \|w\|_1\left(\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) - 8\sqrt{\frac{\ln(2d)+\ln(4/\delta)}{n}}\right) \\
&\geq \|w\|_1 \mathrm{Bal}(\mu_{|\mathcal{D}_\star})/2 > 0.
\end{aligned}$$

Since $\int \max\{Aw,0\}d\widehat{\mu}_{|\mathcal{D}_\star} > 0$, there must exist $z_i\in\mathcal{S}$ with $(Aw)(z_i) > 0$, and in particular $w\notin \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$. To see how this gives the result, suppose contradictorily that $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) \subsetneq \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, whereby there must exist $w\in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star}) \cap \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$ with $w\neq 0$. But the above analysis showed that every $w\in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$ with $\|w\|_1 > 0$ has $w\notin \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, a contradiction. ■

Next lemma is a refined analysis of deviations over $\mathcal{D}_\star$ under the assumption $\ell\in\mathbb{L}_{\mathrm{b}}^{2+}$. In particular, $\mathrm{Bal}(\mu_{|\mathcal{D}_\star})$ will be used to establish strong convexity of $\mathcal{R}(\cdot;\mu_{|\mathcal{D}_\star})$ around $\bar{w}$. The core of the convergence rate argument itself follows almost identically a proof by Shalev-Shwartz et al. (2008, Theorem 1), with two important differences that necessitated a careful reproof.

- Rather than controlling a function which is strongly convex everywhere thanks to a regularizer, it is instead only used that $\mathcal{R}(\cdot;\mu_{|\mathcal{D}_\star})$ is inherently strongly convex around the optimum without any regularization.
- This strong convexity around the optimum is only established over $\mu_{|\mathcal{D}_\star}$, and in particular not over $\widehat{\mu}_{|\mathcal{D}_\star}$. Of course, since $\mathrm{Bal}(\mu_{|\mathcal{D}_\star})$ is statistically stable (see Lemma J.10), the same proof shows that $\mathcal{R}(\cdot;\widehat{\mu}_{|\mathcal{D}_\star})$ is also strongly convex along $\mathcal{D}_\star$, but it is interesting and pleasant that the proof works directly without establishing this.

Lemma J.11 *Let probability measure μ over $\mathcal{Z}$, hypotheses $\mathcal{H}$ with $|\mathcal{H}| < \infty$, loss function $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$, and canonical difficult set $\mathcal{D}_\star$ with $\mu(\mathcal{D}_\star) > 0$ be given. Let a primal-dual optimal pair $(\bar{w}, \bar{q})$ for Eq. (7) with measure $\mu_{|\mathcal{D}_\star}$ be given with $\bar{w} \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$. Lastly, let $B \geq \|\bar{w}\|_1$ be given, and set $\mathcal{W} := \{w \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp : \|w\|_1 \leq B\}$. The following statements hold.*

1. *Set $\tau := \inf_{|z| \leq B} \ell''(z)$ and $\lambda := \tau \mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2$, where $\tau > 0$ since $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$. Then, for every $w \in \mathcal{W}$,*
$$\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) \geq \frac{\lambda}{2}\|w - \bar{w}\|_1^2. \tag{19}$$
2. *Let a draw $\mathcal{S}$ from $\mu_{|\mathcal{D}_\star}$ of size n be given. Then, with probability at least $1-\delta$, every $w \in \mathcal{W}$ satisfies*
$$\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) \leq 2\mathcal{E}(w;\widehat{\mu}_{|\mathcal{D}_\star}) + \frac{1024\ell'(2B)^2(\ln(2d) + \ln(4/\delta))}{\lambda n}.$$

Proof

1. To start, applying Taylor's theorem pointwise, every $w \in \mathcal{W}$ satisfies
$$\begin{aligned}\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) &= \int \ell(Aw)d\mu_{|\mathcal{D}_\star} - \int \ell(A\bar{w})d\mu_{|\mathcal{D}_\star} \\ &\geq \underbrace{\int \ell'(A\bar{w})(Aw - A\bar{w})d\mu_{|\mathcal{D}_\star}}_{\heartsuit} + \underbrace{\frac{1}{2}\int \tau(Aw - A\bar{w})^2 d\mu_{|\mathcal{D}_\star}}_{\triangle}.\end{aligned}$$
To manage $\heartsuit$, since $\bar{q} = \ell'(A\bar{w})$ μ-a.e. (by Theorem 2.1), and since $\bar{q}$ is dual feasible (whereby $A^\top \bar{q} = 0$), then
$$\heartsuit = \int \ell'(A\bar{w})(Aw - A\bar{w})d\mu_{|\mathcal{D}_\star} = \int \bar{q}(A(w - \bar{w}))d\mu_{|\mathcal{D}_\star} = 0.$$
For the second term $\triangle$, by Jensen's inequality and the definition of $\mathrm{Bal}(\mu_{|\mathcal{D}_\star})$,
$$\begin{aligned}\triangle &\geq \frac{\tau}{2}\left(\int |A(w - \bar{w})|d\mu_{|\mathcal{D}_\star}\right)^2 \\ &\geq \frac{\tau}{2}\left(\int \max\{A(w - \bar{w}), 0\}d\mu_{|\mathcal{D}_\star}\right)^2 \\ &\geq \frac{\tau \mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2}{2}\|w - \bar{w}\|_1^2 \ ,\end{aligned}$$
which gives the bound.

2. As discussed above, this proof follows one due to Shalev-Shwartz et al. (2008, Proof of Theorem 1).

   To start, when $\nu \in \{\mu_{|\mathcal{D}_\star}, \widehat{\mu}_{|\mathcal{D}_\star}\}$ and $\ell \in \mathbb{L}_{\mathsf{b}}^{2+}$, then $\mathcal{R}(\cdot;\nu)$ is $L := \ell'(2B)$ Lipschitz; additionally, it satisfies Eq. (19), meaning $\mathcal{R}(\cdot;\nu)$ is λ-strongly-convex around $\bar{w}$ as above.

   Let $r > 0$ be a constant to be optimized at the end of the proof, and define

$$\begin{aligned} k_w &:= \min\left\{k \in \mathbb{Z}_+ : \mathcal{E}(w;\mu_{|\mathcal{D}_\star}) \le r4^k\right\}, \\ f_w(z) &:= \ell\big((Aw)(z)\big) - \ell\big((A\bar{w})(z)\big), \\ g_w(z) &:= 4^{-k_w} f_w(z), \\ \mathcal{G} &:= \{g_w : w \in \mathcal{W}\}. \end{aligned}$$

   Applying Lemma C.1 to $\mathcal{G}$, then with probability at least $1-\delta$, each $w \in \mathcal{W}$ satisfies

$$\int g_w d\mu_{|\mathcal{D}_\star} \le \int g_w d\widehat{\mu}_{|\mathcal{D}_\star} + 2\underbrace{\mathfrak{R}(\mathcal{G})}_{\spadesuit} + 4\underbrace{\sup_{w\in\mathcal{W},z\in\mathcal{Z}} |g_w(z)|}_{\diamondsuit}\sqrt{\frac{2\ln(4/\delta)}{n}}. \tag{20}$$

   Following the proof scheme of Shalev-Shwartz et al. (2008, Proof of Theorem 1), the two critical terms are bounded as follows.

   - First, $\diamondsuit = \sup_{w\in\mathcal{W},z\in\mathcal{Z}} |g_w(z)| \le L\sqrt{2r/\lambda}$ as follows. For any $w \in \mathcal{W}$ and any $z \in \mathcal{Z}$, by the fact that $\mathcal{R}(\cdot;\mu_{|\mathcal{D}_\star})$ is L-Lipschitz and satisfies Eq. (19), since $\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) \le r4^{k_w}$ by definition of $k_w \ge 0$,

$$\begin{aligned} |g_w(z)| &= 4^{-k_w}\left|\ell\big((Aw)(z)\big) - \ell\big((A\bar{w})(z)\big)\right| \\ &\le 4^{-k_w} L\|w-\bar{w}\|_1 \\ &\le 4^{-k_w} L\sqrt{2\mathcal{E}(w;\mu_{|\mathcal{D}_\star})/\lambda} \\ &\le 4^{-k_w} L\sqrt{2r4^{k_w}/\lambda} \\ &= 4^{-k_w/2} L\sqrt{2r/\lambda} \\ &\le L\sqrt{2r/\lambda} \end{aligned}$$

     as desired.

   - Second, $\spadesuit = \mathfrak{R}(\mathcal{G}) \le 4L\sqrt{r\ln(2d)/(\lambda n)}$. For this, first define two helper classes

$$\begin{aligned} \mathcal{F}(a) &:= \left\{f_w : w \in \mathcal{W}, \mathcal{E}(w;\mu_{|\mathcal{D}_\star}) \le a\right\}, \\ \tilde{\mathcal{F}}(a) &:= \left\{f_w : w \in \mathcal{W}, \|w-\bar{w}\|_1 \le \sqrt{2a/\lambda}\right\}. \end{aligned}$$

By Eq. (19), $f_w \in \mathcal{F}(a)$ implies $\|w - \bar{w}\|_1 \leq \sqrt{2a/\lambda}$, thus $\mathcal{F}(a) \subseteq \tilde{\mathcal{F}}(a)$. By various properties of Rademacher complexity from Lemma C.2,

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}(a)) &\leq \mathfrak{R}(\tilde{\mathcal{F}}(a)) \\
&\leq L\mathfrak{R}(\{z \mapsto (Aw)(z) : w \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp, \|w - \bar{w}\|_1 \leq \sqrt{2a/\lambda}\}) \\
&\leq L\mathfrak{R}(\{z \mapsto (Aw)(z) : w \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp, \|w\|_1 \leq \sqrt{2a/\lambda}\}) \\
&\leq L\sqrt{\frac{4a\ln(2d)}{\lambda n}}.
\end{aligned}$$

To control $\mathfrak{R}(\mathcal{G})$ first note that $0 \in \mathcal{G}$ and $0 \in \mathcal{F}(a)$ for any $a \geq 0$, since these sets all consider the choice $\bar{w} \in \mathcal{W}$. Consequently, Lemma C.2.ii may be applied, which together with Lemma C.2.i yields

$$\mathfrak{R}(\mathcal{G}) \leq \mathfrak{R}\left(\cup_{k=0}^{\infty} 4^{-k}\mathcal{F}(r4^k)\right) \leq \sum_{k=0}^{\infty} 4^{-k}\mathfrak{R}\left(\mathcal{F}(r4^k)\right).$$

This completes the bound on $\mathfrak{R}(\mathcal{G})$, since the above estimates grant

$$\sum_{k=0}^{\infty} 4^{-k}\mathfrak{R}\left(\mathcal{F}(r4^k)\right) \leq L\sqrt{\frac{4r\ln(2d)}{\lambda n}} \sum_{k=0}^{\infty} 4^{-k/2} = 4L\sqrt{\frac{r\ln(2d)}{\lambda n}}.$$

Continuing with the deviation bound in Eq. (20), set r with foresight as

$$r := 8192L^2\left(\frac{\ln(2d) + \ln(4/\delta)}{\lambda n}\right).$$

Now combining the preceding inequalities on $\diamondsuit$ and $\spadesuit$, the choice of r, and the general inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ over nonnegative reals, it follows for every $w \in \mathcal{W}$ that

$$\begin{aligned}
\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) - \mathcal{E}(w;\widehat{\mu}_{|\mathcal{D}_\star}) &= \mathcal{R}(w;\mu_{|\mathcal{D}_\star}) - \mathcal{R}(\bar{w};\mu_{|\mathcal{D}_\star}) - \left(\mathcal{R}(w;\widehat{\mu}_{|\mathcal{D}_\star}) - \inf_{v\in\mathbb{R}^d} \mathcal{R}(w;\widehat{\mu}_{|\mathcal{D}_\star})\right) \\
&\leq \mathcal{R}(w;\mu_{|\mathcal{D}_\star}) - \mathcal{R}(\bar{w};\mu_{|\mathcal{D}_\star}) - \left(\mathcal{R}(w;\widehat{\mu}_{|\mathcal{D}_\star}) - \mathcal{R}(\bar{w};\widehat{\mu}_{|\mathcal{D}_\star})\right) \\
&= \int f_w d\mu_{|\mathcal{D}_\star} - \int f_w d\widehat{\mu}_{|\mathcal{D}_\star} \\
&= 4^{k_w}\left(\int g_w d\mu_{|\mathcal{D}_\star} - \int g_w d\widehat{\mu}_{|\mathcal{D}_\star}\right) \\
&\leq 4^{k_w}\left(8L\sqrt{\frac{r\ln(2d)}{\lambda n}} + 4L\sqrt{\frac{4r\ln(4/\delta)}{\lambda n}}\right) \\
&\leq 4^{k_w} \cdot \sqrt{r} \cdot 8L\sqrt{2} \cdot \sqrt{\frac{\ln(2d) + \ln(4/\delta)}{\lambda n}} \\
&= \frac{r4^{k_w}}{8}.
\end{aligned}$$

To finish the proof, consider two cases for the value of k_w: either $k_w = 0$, or $k_w > 0$. When $k_w = 0$, then the choice of r gives

$$\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) - \mathcal{E}(w;\widehat{\mu}_{|\mathcal{D}_\star}) \leq \frac{r4^0}{8} = \frac{1024L^2(\ln(2d) + \ln(4/\delta))}{\lambda n},$$

which yields the desired inequality since $L \leq \ell'(B)$ and by adding $\mathcal{E}(w; \widehat{\mu}_{|\mathcal{D}_\star}) \geq 0$ to the right hand side. On the other hand, when $k_w > 0$, then the definition of k_w implies $\mathcal{E}(w; \mu_{|\mathcal{D}_\star}) > 4^{k_w - 1} r$, which plugged back into the above gives

$$\mathcal{E}(w; \mu_{|\mathcal{D}_\star}) - \mathcal{E}(w; \widehat{\mu}_{|\mathcal{D}_\star}) \leq \frac{r 4^{k_w}}{8} \leq \frac{r 4^{k_w - 1}}{2} < \frac{1}{2}\mathcal{E}(w; \mu_{|\mathcal{D}_\star}),$$

meaning $\mathcal{E}(w; \mu_{|\mathcal{D}_\star}) \leq 2\mathcal{E}(w; \widehat{\mu}_{|\mathcal{D}_\star})$, giving the desired bound.

■

Proof (of Lemma 3.10) This proof will be focused on parts (ii) and (iii); to start, note the following two supporting results, the second of which will establish part (i) of the desired statement along the way.

- First note how $\inf_{w \in \mathbb{R}^d} \mathcal{R}(w; \nu)$ can be related for $\nu \in \{\mu_{|\mathcal{D}_\star}, \widehat{\mu}_{|\mathcal{D}_\star}\}$. By Proposition J.4 and the assumption $\mu(\mathcal{D}_\star) > 0$, $\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) > 0$, and thus Lemma J.5 gives a primal optimum $\bar{w} \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$ with $\|\bar{w}\|_1 \leq \ell(0)/(\bar{s}\mathrm{Bal}(\mu_{|\mathcal{D}_\star}))$. Consequently, by Hoeffding's inequality applied to a random variable with range $\ell(\|\bar{w}\|_1)$, with probability at least $1 - \delta$,

$$\begin{aligned}\inf_{v \in \mathbb{R}^d} \mathcal{R}(v; d\widehat{\mu}_{|\mathcal{D}_\star}) \leq \mathcal{R}(\bar{w}; d\widehat{\mu}_{|\mathcal{D}_\star}) &\leq \mathcal{R}(\bar{w}; d\mu_{|\mathcal{D}_\star}) + \ell(\|\bar{w}\|_1)\sqrt{\frac{2\ln(1/\delta)}{n}}. \\ &= \inf_{v \in \mathbb{R}^d} \mathcal{R}(v; d\mu_{|\mathcal{D}_\star}) + \ell(\|\bar{w}\|_1)\sqrt{\frac{2\ln(1/\delta)}{n}},\end{aligned}$$

  which will be useful via the rearrangement

$$-\inf_{v \in \mathbb{R}^d} \mathcal{R}(v; d\mu_{|\mathcal{D}_\star}) \leq -\inf_{v \in \mathbb{R}^d} \mathcal{R}(v; d\widehat{\mu}_{|\mathcal{D}_\star}) + \ell(\|\bar{w}\|_1)\sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{21}$$

- Secondly, assume the final consequence of Lemma J.10 holds, discarding along the way another failure event having probability at most δ: by the lower bound on n, $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) = \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$ and

$$2\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star}) \geq \mathrm{Bal}(\mu_{|\mathcal{D}_\star}). \tag{22}$$

  To see the value of $\mathrm{Ker}(\mu_{|\mathcal{D}_\star}) = \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$, given any $w \in \mathbb{R}^d$, henceforth write $w = w_0 \oplus w_\perp$ with $w_0 \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})$ and $w_\perp \in \mathrm{Ker}(\mu_{|\mathcal{D}_\star})^\perp$, where additionally $w_0 \in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})$ and $w_\perp \in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$. As a first consequence,

$$(Aw)(z) = (Aw_0)(z) + (Aw_\perp)(z) = (Aw_\perp)(z) \quad \text{for } \mu\text{-a.e. and } \widehat{\mu}\text{-a.e. } z \in \mathcal{D}_\star. \tag{23}$$

  which further implies

$$\mathcal{R}(w_\perp; \nu) = \mathcal{R}(w; \nu) \qquad \text{and} \qquad \mathcal{E}(w_\perp; \nu) = \mathcal{E}(w; \nu) \qquad \text{for } \nu \in \{\mu_{|\mathcal{D}_\star}, \widehat{\mu}_{|\mathcal{D}_\star}\}. \tag{24}$$

Secondly, by Lemma J.1 applied to measure $\widehat{\mu}_{|\mathcal{D}_\star}$ (where Lemma J.1 requires $w_\perp \in \mathrm{Ker}(\widehat{\mu}_{|\mathcal{D}_\star})^\perp$), and also using Eq. (22), Eq. (24), and the form of B_w,

$$\|w_\perp\|_1 \le \frac{\mathcal{R}(w_\perp; \widehat{\mu}_{|\mathcal{D}_\star})}{\bar{s}\mathrm{Bal}(\widehat{\mu}_{|\mathcal{D}_\star})} \le \frac{2\mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star})}{\bar{s}\mathrm{Bal}(\mu_{|\mathcal{D}_\star})} \le B_w. \tag{25}$$

This last inequality is essential as it allows $\|w_\perp\|_1$ and $\mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star})$ to be related, the latter being a purely sample-dependent quantity. In particular, part (i) follows immediately by combining Eq. (23) and Eq. (25); that is, for μ-a.e. $z \in \mathcal{D}_\star$ and $\widehat{\mu}$-a.e. $z \in \mathcal{D}_\star$, $|(Aw)(z)| = |(Aw_\perp)(z)| \le \|w_\perp\|_1 \le B_w$.

The remainder of the proof will establish parts (ii) and (iii) by organizing $\mathbb{R}^d$ into sets $(\mathcal{W}_i)_{i\ge 1}$ with $\mathbb{R}^d = \cup_{i\ge 1}\mathcal{W}_i$. For every integer $i \ge 1$ define

$$\begin{aligned} R_i &:= i + \|\bar{w}\|_1, \\ \mathcal{W}_i &:= \left\{ w \in \mathbb{R}^d : \|w_\perp\|_1 \le R_i \right\}, \\ \delta_i &:= \delta/(i+1)^2. \end{aligned}$$

By this choice, $\sum_{i\ge 1}\delta_i \le \delta$, and thus proving both types of bound contributes to the final 2δ in the full statement. Secondly, note how Eq. (25) gives a way to use $\mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star})$ to choose i with $w \in \mathcal{W}_i$: the largest i granting $w \in \mathcal{W}_i$ satisfies $i \le 1 - \|\bar{w}\|_1 + \|w_\perp\|_1 \le B_w - 1 - \|\bar{w}\|_1$.

With this structure in place, parts (ii) and (iii) are established for each $i \in \mathbb{Z}_{++}$ separately as follows, and the general bounds follow by replacing the term R_i via

$$R_i \le i + \|\bar{w}\|_1 \le \left(B_w - 1 - \|\bar{w}\|_1\right) + \|\bar{w}\|_1 = B_w - 1.$$

Note that, restricted to $\mathcal{W}_i$, ℓ satisfies μ-a.e. boundededness in the sense that $\sup_{w\in\mathcal{W}_i} \ell((Aw)(z)) \le \ell(R_i)$ for μ-a.e. $z \in \mathcal{D}_\star$ and also $\widehat{\mu}$-a.e. $z \in \mathcal{D}_\star$ (by part (i)), and ℓ is Lipschitz with constant $\ell'(R_i)$.

(ii) Using Rademacher complexity of Lipschitz functions (Lemmas C.1 and C.2), and Eq. (24) to swap w and $w_\perp$, and noting the general inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2a + 2b}$ for nonnegative reals, then for any fixed i, with probability at least $1 - \delta_i$, every $w \in \mathcal{W}_i$ satisfies

$$\begin{aligned} \mathcal{R}(w; \mu_{|\mathcal{D}_\star}) - \mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star}) &= \mathcal{R}(w_\perp; \mu_{|\mathcal{D}_\star}) - \mathcal{R}(w_\perp; \widehat{\mu}_{|\mathcal{D}_\star}) \\ &\le 2\ell'(R_i)R_i\sqrt{2\ln(2d)/n} + 4\ell(R_i)\sqrt{2\ln(4/\delta_i)/n} \\ &\le 8(\ell'(R_i)R_i + \ell(R_i))\sqrt{\frac{\ln(2d) + \ln(4/\delta_i)}{n}} \\ &\le 8\ell(2R_i)\sqrt{\frac{\ln(2d) + \ln(4/\delta_i)}{n}}, \end{aligned}$$

where the last simplification used $\ell(R_i) + \ell'(R_i)(2R_i - R_i) \le \ell(2R_i)$ by convexity. To finish the proof, combining the above display with Eq. (21) gives

$$\begin{aligned} \mathcal{E}(w; \mu_{|\mathcal{D}_\star}) - \mathcal{E}(w; \widehat{\mu}_{|\mathcal{D}_\star}) &= \mathcal{R}(w; \mu_{|\mathcal{D}_\star}) - \mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star}) - \inf_{v\in\mathbb{R}^d} \mathcal{R}(v; \mu_{|\mathcal{D}_\star}) + \inf_{v\in\mathbb{R}^d} \mathcal{R}(v; \widehat{\mu}_{|\mathcal{D}_\star}) \\ &\le \mathcal{R}(w; \mu_{|\mathcal{D}_\star}) - \mathcal{R}(w; \widehat{\mu}_{|\mathcal{D}_\star}) + \ell(\|\bar{w}\|_1)\sqrt{\frac{2\ln(1/\delta)}{n}}. \\ &\le 10\ell(2R_i)\sqrt{\frac{\ln(2d) + \ln(4/\delta_i)}{n}}. \end{aligned}$$

(iii) Similarly to the purely Lipschitz case above, but now using Lemma J.11 to control deviations, with probability at least $1-\delta_i$, each $w \in \mathcal{W}_i$ satisfies

$$\begin{aligned}\mathcal{E}(w;\mu_{|\mathcal{D}_\star}) &= \mathcal{E}(w_\perp, \mu_{|\mathcal{D}_\star}) \\ &\leq 2\mathcal{E}(w_\perp;\widehat{\mu}_{|\mathcal{D}_\star}) + \frac{1024\ell'(2R_i)^2(\ln(2d)+\ln(4/\delta_i))}{n\tau(R_i)\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2} \\ &= 2\mathcal{E}(w;\widehat{\mu}_{|\mathcal{D}_\star}) + \frac{1024\ell'(2R_i)^2(\ln(2d)+\ln(4/\delta_i))}{n\tau(R_i)\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2}.\end{aligned}$$

■

J.5. Proof of Theorem 1.2

Before proving Theorem 1.2, note briefly how samples drawn from μ can be treated as a draw from $\mu_{|\mathcal{D}_\star}$ and $\mu_{|\mathcal{D}_\star^c}$.

Lemma J.12 *Let probability measure μ and a canonical difficult set $\mathcal{D}_\star$ be given. Let $\mathcal{S}$ denote a draw from μ of size $n \geq 8\ln(1/\delta)$, and define $\mathcal{S}_\mathcal{D} := \mathcal{S}\cap\mathcal{D}_\star$ and $\mathcal{S}_c := \mathcal{S}\cap\mathcal{D}_\star^c$ with sizes $n_\mathcal{D} := |\mathcal{S}_\mathcal{D}|$ and $n_c := |\mathcal{S}_c|$. Then with probability at least $1-\delta$ over the draw of $\mathcal{S}$,*

$$n_\mathcal{D} \geq n\mu(\mathcal{D}_\star)/2, \qquad\qquad n_c \geq n\mu(\mathcal{D}_\star^c)/2,$$

and $\mathcal{S}_\mathcal{D}$ and $\mathcal{S}_c$ can be treated as draws of size $n_{\mathcal{D}_\star}$ and n_c from $\mu_{|\mathcal{D}_\star}$ and $\mu_{|\mathcal{D}_\star^c}$, respectively.

Proof Treating the partitioned sample as two independent draws is the usual rejection sampling. Moreover, by multiplicative Chernoff bounds (Kearns and Vazirani, 1994, Theorem 9.2) and the lower bound on n,

$$\begin{aligned}n_{\mathcal{D}_\star}/n = \widehat{\mu}(\mathcal{D}_\star) &\geq \mu(\mathcal{D}_\star)\left(1-\sqrt{2\ln(1/\delta)/n}\right) \geq \mu(\mathcal{D}_\star)/2, \\ n_c/n = \widehat{\mu}(\mathcal{D}_\star^c) &\geq \mu(\mathcal{D}_\star^c)\left(1-\sqrt{2\ln(1/\delta)/n}\right) \geq \mu(\mathcal{D}_\star^c)/2.\end{aligned}$$

■

All the pieces are in place to prove Theorem 1.2.

Proof (of Theorem 1.2) To prove the bound, set $\delta' := \delta/7$, and let a sample $\mathcal{S}$ be given with size

$$n \geq 8\ln(1/\delta') + \mathbf{1}[\mu(\mathcal{D}_\star) > 0]\left(\frac{512(\ln(2d)+\ln(4/\delta'))}{\mu(\mathcal{D}_\star)\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2}\right) = \Omega(\ln(1/\delta')).$$

By Lemma J.12, conditioning away a first failure probability of δ', $\mu(\mathcal{D}_\star) > 0$ implies the set $\mathcal{S}_\mathcal{D} = \mathcal{S}\cap\mathcal{D}_\star$ is an i.i.d. draw from $\mu_{|\mathcal{D}_\star}$ of size $n_\mathcal{D} = \Omega(n)$ satisfying moreover

$$n_\mathcal{D} \geq \frac{256(\ln(2d)+\ln(4/\delta'))}{\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2}, \tag{26}$$

whereas $\mu(\mathcal{D}_\star^c) > 0$ implies the set $\mathcal{S}_c = \mathcal{S} \cap \mathcal{D}_\star^c$ is an i.i.d. draw from $\mu_{|\mathcal{D}_\star^c}$ of size $n_c = \Omega(n)$.

Let $w \in \mathbb{R}^d$ be arbitrary, and note

$$\int |\eta_w - \bar{\eta}| d\mu = \underbrace{\int_{\mathcal{D}_\star} |\eta_w - \bar{\eta}| d\mu}_{\heartsuit} + \underbrace{\int_{\mathcal{D}_\star^c} |\eta_w - \bar{\eta}| d\mu}_{\triangle}; \tag{27}$$

the proof will proceed by controlling $\heartsuit$ and $\triangle$ separately, where either term is 0 automatically if either $\mu(\mathcal{D}_\star) = 0$ or $\mu(\mathcal{D}_\star^c) = 0$, respectively. Note, throughout this proof, that $\mathcal{D} = \mathcal{D}_\star$ μ-a.e. since $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$ thanks to Proposition 3.9.

First consider the term $\heartsuit$ (when $\mu(\mathcal{D}_\star) > 0$); the goal will be to invoke Lemma 3.6, however many of the messy terms therein will be controlled via Lemma 3.10. In particular, assume the various parts of Lemma 3.10, and condition away an additional $4\delta'$ failure probability, noting that $n_{\mathcal{D}}$ is sufficiently large by Eq. (26). Let B_w be as in the statement of Lemma 3.10, which crucially only depends on w only through $\hat{\mathcal{E}}_n(w)$, and satisfies $|(Aw)(z)| \le B_w$ for μ-a.e. $z \in \mathcal{D}_\star$. Furthermore, since $\mathrm{Bal}(\mu_{|\mathcal{D}_\star}) > 0$ by $\mu(\mathcal{D}_\star) > 0$ and Lemma J.3, by Lemma J.5 there exists a primal optimum $\bar{w}$ with $\bar{q}_{\mathcal{D}_\star} \in \partial\ell(A\bar{w})$ μ-a.e. over $\mathcal{D}_\star$, and μ-a.e. $z \in \mathcal{D}_\star$ satisfies

$$|(A\bar{w})(z)| \le \|\bar{w}\|_1 \le \frac{\ell(0)\mu_{|\mathcal{D}_\star}(\mathcal{Z})}{\ell'(0)\mathrm{Bal}(\mu_{|\mathcal{D}_\star})} \le B_w.$$

By Lemma 3.2, $\tilde{q}(z) := \bar{q}_{\mathcal{D}_\star}(z)\mathbf{1}[z \in \mathcal{D}_\star]$ is also optimal for the full problem over μ; but Theorem 2.1 provided that the full dual optimum is μ-a.e. unique, meaning the general $\bar{q}$ and this specialized $\tilde{q}$ agree μ-a.e. over $\mathcal{D}_\star$, and in particular μ-a.e. $z \in \mathcal{D}_\star$ satisfies

$$\ell'(-B_w) \le \ell'((A\bar{w})(z)) \quad = \quad \bar{q}(z) \quad = \quad \ell'((A\bar{w})(z)) \le \ell'(B_w).$$

Consequently, applying Lemma 3.6 with constants $c_1 := B_w$, $c_2 := \inf_{|r| \le B_w} \ell'(r) = \ell'(-B_w)$, and $c_3 := \sup_{|r| \le B_w} \ell'(r) = \ell'(B_w)$, we have

$$\mu(S_-) = \mu(S_+) = \mu(V) = 0 \ ,$$

so it suffices to include the term for U. Now using Lemma 3.10.iii to relate $\mathcal{E}(w; \mu_{|\mathcal{D}_\star})$ and $\mathcal{E}_n(w; \widehat{\mu}_{|\mathcal{D}_\star})$, additionally the general inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for nonnegative reals, and lastly recalling the notation $\tau(B_w) := \inf_{|q| \le B_w} \ell''(q)$ from Lemma 3.10,

$$\begin{aligned}
\heartsuit &\le L_\phi \sqrt{\frac{2\mathcal{E}(w; \mu_{\mathcal{D}_\star})}{\tau(B_w)}} \\
&\le L_\phi \sqrt{\frac{2\mu(\mathcal{D}_\star)}{\tau(B_w)}} \left(\sqrt{2\hat{\mathcal{E}}_n(w; \mu_{|\mathcal{D}_\star})} + \sqrt{\frac{1024\ell'(2B_w)^2(\ln(2d) + \ln(4B_w^2/\delta'))}{n_{\mathcal{D}}\mathrm{Bal}(\mu_{|\mathcal{D}_\star})^2\tau(B_w)}} \right) \\
&\le \mathcal{O}\left(f_2\left(\hat{\mathcal{E}}_n(w)\right) \left(\sqrt{\hat{\mathcal{E}}_n(w)} + \sqrt{\frac{\ln(1/\delta)}{n}} \right) \right),
\end{aligned}$$

where the term $f_2(\mathcal{E}_n(w))$ collects all terms depending on B_w, which itself depends on w only through $\hat{\mathcal{E}}_n(w)$ as per Lemma 3.10.

Now consider the term $\triangle$ (when $\mu(\mathcal{D}_\star^c) > 0$); the goal will be to invoke Lemma 3.5, however once again some terms in the bound will be handled manually via Lemma J.9. Set $\varepsilon := r := \sqrt{\hat{\mathcal{E}}_n(w)}$, and define $S_r := \{z \in \mathcal{D}_\star^c : \ell\big((Aw)(z)\big) \geq r\}$ exactly as in Lemma 3.5, and which also appears in Lemma J.9; applying Lemma J.9 with this ε to w (where $\varepsilon > 0$ since $\mathcal{E}(w; d\widehat{\mu}_{\mathcal{D}_\star}) > 0$ by $\mu(\mathcal{D}_\star) > 0$ and the assumed lower bound on n_c and since $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$), and discarding an additional $2\delta'$ failure probability along the way, $\mu(S_r) \leq \mathcal{O}\Big(\hat{\mathcal{E}}_n(w)^{1/2} + \sqrt{(\ln(n_c) + \ln(1/\delta'))/n_c}\Big)$. Combining this bound on $\mu(S_r)$ with the bound on $|\eta_w - \bar{\eta}|$ from Lemma 3.5 (which uses the fact that $\mathcal{D} = \mathcal{D}_\star$ μ-a.e. since $\ell \in \mathbb{L}_{\mathrm{b}}^{2+}$) gives

$$\triangle \leq \mu(S_r) + r\mu(\mathcal{D}_\star^c \setminus S_r) \max\left\{\frac{1}{\ell(0)}, \frac{c_\ell}{\ell'(0)}\right\} = \mathcal{O}\left(\sqrt{\hat{\mathcal{E}}_n(w)} + \sqrt{\frac{\ln(n) + \ln(1/\delta)}{n}}\right).$$

Plugging these bounds on $\triangle$ and $\heartsuit$ back into Eq. (27) gives the desired inequality.

Lastly, the convergence statement is, as usual, a consequence of the Borel-Cantelli lemma. In particular, let $\varepsilon > 0$ be arbitrary, set $\delta_n := 1/n^2$, and define the event

$$E_{n,\varepsilon} := \left[\int |\eta_{w_n} - \bar{\eta}| d\mu > \varepsilon\right].$$

Applying the bound above for each w_n, as $n \to \infty$ and $\hat{\mathcal{E}}_n(w_n) \to 0$, we obtain that there exists some N so that every $n > N$ has $\Pr(E_{n,\varepsilon}) \leq \delta_n$. Consequently,

$$\sum_{n \geq 1} \Pr(E_{n,\varepsilon}) \leq \sum_{n=1}^{N} 1 + \sum_{n > N} \delta_n < \infty,$$

and the result follows by applying the Borel-Cantelli lemma. ■