On Misspecification in Prediction Problems and Robustness via Improper Learning

Annie Marsden * John Duchi † Gregory Valiant ‡

Stanford University

Abstract

We study probabilistic prediction games when the underlying model is misspecified, investigating the consequences of predicting using an incorrect parametric model. We show that for a broad class of loss functions and parametric families of distributions, the regret of playing a "proper" predictor—one from the putative model class—relative to the best predictor in the same model class has lower bound scaling at least as $\sqrt{\gamma n}$, where γ is a measure of the model misspecification to the true distribution in terms of total variation distance. In contrast, using an aggregation-based (improper) learner, one can obtain regret $d \log n$ for any underlying generating distribution, where d is the dimension of the parameter; we exhibit instances in which this is unimprovable even over the family of all learners that may play distributions in the convex hull of the parametric family. These results suggest that simple strategies for aggregating multiple learners together should be more robust, and several experiments conform to this hypothesis.

1 Introduction

Suppose we seek a probability distribution $p(y \mid x)$ modeling outcomes y given data x. The typical approach is to choose a parametric family of probability distributions, then find the "best" member of this family according to a given loss. It is rarely realistic to assume that the parametric family is well-specified, and thus it is important to understand the consequences of misspecification and how to circumvent these downsides. To address these challenges, in this paper we derive a new measure of a problem's robustness to misspecification that relies on the curvature of the loss at hand and putative parametric family, proving that this measure lower bounds convergence rates for prediction error and certifies the *failure* of a parametric family and loss to be robust (or achieve optimal convergence rates for prediction). To complement this new family of lower bounds for probabilistic prediction problems, we build out of earlier work on *improper learning* [40, 14]—when we may choose predictions $p(y \mid x)$ outside the given model family—to show how it is possible to be robust to such misspecification, and moreover, we give new optimality guarantees for such improper procedures.

Formalizing our setting, we consider the following probabilistic game: a player receives a covariate vector $x \in \mathcal{X}$, plays a distribution $p(\cdot \mid x)$ on a target set $\mathcal{Y}$, then receives $y \in \mathcal{Y}$ and suffers loss

$$L(p(\cdot \mid x), y).$$

We study both a sequential and a stochastic variant of this problem. In the former, for a sequence of examples $\{(x_i, y_i)\}_{i=1}^n$, a player chooses a distribution p_k depending on the past examples $\{(x_i, y_i)\}_{i=1}^{k-1}$,

*Supported by ACM SIGHPC/Intel Fellowship

†Supported by NSF CAREER Award CCF-1553086, ONR YIP N00014-19-2288, and the DAWN Consortium

‡Supported by DOE award DE-SC0019205, NSF awards 1813049 and 1704417, and ONR YIP N00014-18-1-2295

and then for a fixed conditional distribution p on $Y \mid X$, suffers regret

$$\mathsf{Reg}_n(p) := \sum_{i=1}^n L(p_i(\cdot \mid x_i), y_i) - \sum_{i=1}^n L(p(\cdot \mid x_i), y_i).$$

In the stochastic variant, the examples (x_i, y_i) are i.i.d. from an unknown distribution P, and we consider the risk of the conditional p.m.f. p,

$$\mathsf{Risk}_P(p) := \mathbb{E}[L(p(Y \mid X))] = \int L(p(\cdot \mid x), y) dP(x, y),$$

where the expectation is taken over $(X, Y) \sim P$. The goal is to play p_i or p above to make the regret and risk as small as possible.

This regret and risk minimization framework is familiar from the universal prediction and probabilistic forecasting literature [32, 18, 16, 9, 5], which considers best possible estimators and online learners for the regret Reg_n over various losses L and families $\mathcal{P}$ of possible predictive distributions. In this paper, we study these regret and risk minimization formulations over parametric families of distributions $\{p_\theta(\cdot \mid x)\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}^d$ is a convex set. We shall either consider the regret

$$\mathsf{Reg}_n^\Theta := \sup_{\theta^\star \in \Theta} \mathsf{Reg}_n(p_{\theta^\star}) = \sum_{i=1}^n L(p_i(\cdot \mid x_i), y_i) - \inf_{\theta^\star \in \Theta} \sum_{i=1}^n L(p_i(\cdot \mid x_i), y_i) \tag{1a}$$

or—in the stochastic version of the problem—the excess risk relative to this family,

$$\mathsf{Risk}_P^\Theta(p) := \mathsf{Risk}_P(p) - \inf_{\theta^\star \in \Theta} \mathsf{Risk}_P(p_{\theta^\star}). \tag{1b}$$

When the conditional distribution of $Y \mid X$ belongs to the parametric family $\{p_\theta\}_{\theta \in \Theta}$ where $\Theta \subset \mathbb{R}^d$, maximum likelihood estimators enjoy rates of convergence of $O(d/n)$ for the excess risk (1b) as n grows [36]. In typical practice, however, the data generating distribution is misspecified, so it is important to understand how this misspecification impacts possible convergence rates and optimal estimators.

We thus consider three intertwined objects: the parametric family $\{p_\theta\}_{\theta \in \Theta}$ against which we compare the performance of our prediction p, a family Γ of distributions on Y given X that we may play (i.e. predict from), and the family $\mathcal{P}$ of data generating distributions that nature may choose. We study the interaction between these three and the impact of allowing the family $\mathcal{P}$ to differ from the parametric model $\{p_\theta\}_{\theta \in \Theta}$. The traditional approach considers the minimax excess risk over the family Θ,

$$\inf_{\widehat{p}_n} \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta^n} \left[\mathsf{Risk}_P^\Theta(\widehat{p}_n) \right], \tag{2}$$

where the infimum is over all estimators $\widehat{p}_n$ that use the n points $\{(X_i, Y_i)\}_{i=1}^n$ to output a distribution $p(Y \mid X)$, and the expectation is taken over $\{(X_i, Y_i)\}_{i=1}^n \stackrel{\text{iid}}{\sim} P_\theta$, where we have abused notation to use P_θ to denote the joint over (X, Y) when $Y \mid X = x$ follows $p_\theta(\cdot \mid x)$. We elaborate this setting slightly. First, we restrict the estimator $\widehat{p}_n$ to take values in a set Γ of distributions (for example, we might take $\Gamma = \{p_\theta\}_{\theta \in \Theta}$, the parametric family, or its convex hull), which we write as $\widehat{p}_n \in \Gamma$. Second, we expand the supremum (2) to also include distributions P *near* the model P_θ: recalling the definition of the total-variation distance $\|P - Q\|_{\text{TV}} := \sup_A |P(A) - Q(A)|$, we consider distributions P for which there is some $\theta \in \Theta$ such that $\|P - P_\theta\|_{\text{TV}} \le \gamma$. This gives us our misspecified minimax risk.

Definition 1.1. *Let $\Theta \subset \mathbb{R}^d$, $\gamma \ge 0$, and Γ be a set of allowable distributions $p(Y \mid X)$. The* minimax risk *at variation distance γ is*

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) := \inf_{\widehat{p}_n \in \Gamma} \sup_{\theta \in \Theta} \sup_{P : \|P - P_\theta\|_{\text{TV}} \le \gamma} \mathbb{E}_{P^n}[\mathsf{Risk}_P^\Theta(\widehat{p}_n)]. \tag{3}$$

The quantity (3) is somewhat complex. The idea is to quantify—via the parameter γ—the impact of allowing the data generating distribution P to depart slightly from the parametric family $\{p_\theta\}_{\theta\in\Theta}$ while constraining ourselves to play a prediction from the family Γ.

The typical setting in online convex optimization and learning [5] is to take the family of "playable" distributions to be the parametric family $\Gamma = \{p_\theta\}_{\theta\in\Theta}$. In this case, standard minimax risk bounds show that in the well-specified setting that the data comes from the parametric family (i.e. $\gamma = 0$ in Def. 1.1) and the loss L is smooth, then we expect the risk to scale as d/n (cf. [36, 41, 3]). Yet as we show in the first part of this work, such results need not be stable to perturbations away from the parametric model. We show that the curvature of the loss relative to predictions and the parameter space Θ essentially governs convergence rates: when losses are appropriately "flat," there is little information and rates are necessarilty slow and misspecification carries a potentially heavy penalty; conversely, when there is substantial curvature, rates exhibit less antagonistic behavior. Accordingly, we introduce what we term the *linearity constant* Lin of the loss L, family $\{p_\theta\}_{\theta\in\Theta}$, and misspecification γ, showing a lower bound of roughly $\min\{1/\sqrt{n}, \mathrm{Lin}/n\}$ on the minimax risk (3). In some cases we delineate, Lin may be exponentially large in problem parameters, so convergence rates slow to the worst-case rates for general online convex optimization [44, 34, 1], and we consider the family sensitive to misspecification.

To complement these negative results, we highlight a solution to this instability by considering the convex hull of the parameteric family, that is the set of mixtures, aggregations, or ensembles $\mathrm{Conv}\{p_\theta\}_{\theta\in\Theta} := \{\int_{\theta\in\Theta} p_\theta d\mu(\theta) \text{ s.t. } \int_{\theta\in\Theta} d\mu(\theta) = 1 \text{ and } d\mu \ge 0\}$. The idea to combine probabilistic forecasts is classical [17, 15, 26, 7, 19]. When the loss function is *mixable* (which we define later), Vovk's Aggregating Algorithm and its variants, e.g. exponential weights, Exp3, and Bayesian universal prediction [38, 39, 40, 18, 5, 2], provide stability and achieve minimax regret $O(d\log n)$ for any γ in Definition 1.1. By a standard online-to-batch conversion (Jensen's inequality) [6], this guarantees a minimax excess risk (3) of at most $O(d\log n/n)$. We also show that for generalized linear models, Vovk's Aggregating Algorithm is optimal up to log-factors with respect to the misspecification parameter γ. That is, there is no better algorithm to use if you are guaranteed certain values of γ.

It is perhaps unfair to consider the entire convex hull of $\{p_\theta\}_{\theta\in\Theta}$ for the class Γ, as this could potentially yield much smaller risk than the parametric family in the risk (1b). Indeed, we give an example in which the best parameteric predictor has no predictive power, while returning a mixture of two parameterized distributions achieves zero loss (though we also give examples in parametric families where considering the convex hull provides no benefit). To justify the aggregation strategy we give a Bernstein von-Mises theorem under misspecification, which shows that the strategy converges to a Gaussian centered at the risk-minimizing parameter estimate $\widehat{\theta}_n$ with covariance shrinking at rate $O(1/n)$; a corollary of this is that Vovk's Aggregating Algorithm returns a distribution which converges in total variation distance to $p_{\widehat{\theta}_n} \in \{p_\theta\}_{\theta\in\Theta}$. Thus, aggregation (or exponential weights) stabilizes predictions while asymptotically enjoying identical convergence to standard risk-minimization procedures.

1.1 Related Work

Our results broadly fall under probabilistic universal prediction in which the data (x_t, y_t) can be any arbitrary sequence [33, 32, 4, 5, 18, 8]. That Vovk's Aggregating Algorithm provides minimax rate stability is known [14], and this is similar to the minimax guarantees of Bayesian models in universal prediction [18]; a long line of work gives the same logarithmic minimax rates [32, 11, 21, 42]. Early work in these prediction problems focuses on the logarithmic loss $L_{\log}(p(\cdot), y) = -\log p(y)$, while more recent work extends these bounds to exp-concave and so-called "mixable" losses [23, 5]. Our results on minimax lower bounds, distinguishing carefully between well-specified and misspecified models and proper and improper predictions, are novel.

While our results are general, applying to exponential families and beyond, related results are available for logistic regression. In this case, for $B, R > 0$ we consider $\Theta = \{\theta : \|\theta\| \le B\}$, $\mathcal{X} \subset \{x : \|x\| \le R\}$,

and let $\{p_\theta\}_{\theta\in\Theta}$ be the family of binary logistic distributions, $p_\theta(y \mid x) = (1 + e^{-y\theta^T x})^{-1}$, with log loss. Hazan et al. [24] show that any algorithm returning some p_θ suffers minimax risk (recall (3)) $\Omega(\sqrt{B/n})$ in the regime where $n = O(\exp cB)$ for some positive constant $c > 0$, $R = 1$, and the allowable perturbation $\gamma = 1$. Foster et al. [14] show that Vovk's Aggregating Algorithm [40] guarantees minimax risk $O(d \log(Bn)/n)$, allowing one to sidestep this lower bound via improper learning, which we also leverage. In the special case of logistic regression—see Example 1 to come in Section 2.2—a simplification of our results gives lower bound $\Omega(1)\sqrt{\gamma BR/n}$ if $n \le \exp(RB/2)$ and $\Omega(1)\exp(2BR/5)/n$ otherwise. We thus show that even when the perturbations away from the parametric family are small, the minimax risk when the set of playable distributions is $\Gamma = \{p_\theta\}_{\theta\in\Theta}$ may grow substantially; this generalizes Hazan et al. [24], where $R = 1$ and $\gamma = 1$, and gives somewhat sharper constants.

2 Parametric Model Instability

Our first step towards understanding sensitivity to misspecification is to provide optimality guarantees for the minimax risk in Definition 1.1 when the player can play only elements of the parametric family of interest, that is, when $\Gamma = \{p_\theta\}_{\theta\in\Theta}$. We focus on losses that depend specifically on $\theta^T x$, where we have

$$L(p_\theta(\cdot \mid x), y) = \ell(\theta^T x, y) \tag{4}$$

for some twice differentiable and convex $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$. A broad range of models and losses take this form, including all generalized linear models [20].

2.1 Main lower bound

Our key contribution is to lower bound the minimax risk via a quantity we term the *linearity constant* of the induced loss ℓ, which measures the sensitivity of ℓ around various points in its domain. The first component is (roughly) a measure of the signal contained in ℓ for different targets y, where for $t \in \mathbb{R}$, $y \in \mathcal{Y}$, and $\ell'(t_0, y_0)$ as shorthand for $\frac{\partial}{\partial t}\ell(t, y)|_{(t,y)=(t_0,y_0)}$ we define

$$q_\ell^\star(t, y) := \sup_{y_0\in\mathcal{Y},\alpha\in[-1,1]} \left\{ \frac{\alpha\ell'(\alpha t, y_0)}{\alpha\ell'(\alpha t, y_0) - \ell'(t, y)} \;\mid\; \operatorname{sign}(\alpha\ell'(\alpha t, y_0)) \ne \operatorname{sign}(\ell'(t, y)) \right\}. \tag{5}$$

We always have $q_\ell^\star(t, y) \in [0, 1]$. For many cases, this quantity is a positive numerical constant (e.g. for the squared error $\ell(t, y) = \frac{1}{2}(t - y)^2$ with $\mathcal{Y} = \mathbb{R}$, we have $q_\ell^\star(t, y) = 1$). Then for given radii R and B, misspecification size $\gamma \in [0, 1]$, and sample size n, we define

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) := \sup_{\substack{y\in\mathcal{Y} \\ t^2+\delta^2\le\frac{R^2B^2}{2}}} \left\{ |\ell'(t, y)| \min\left\{ \delta\sqrt{\gamma q_\ell^\star(t, y)}, \frac{|\ell'(t, y)|}{\sup_{|\Delta|\le\delta} \ell''(t + \Delta, y)} \frac{q_\ell^\star(t, y)}{\sqrt{2n}} \right\} \right\}. \tag{6}$$

This linearity constant roughly measures the extent to which the loss grows quickly without substantial curvature, that is, $\ell'(t, y)$ is large while $\ell''(t, y)$ is small. A heuristic simplification may help with intuition: by ignoring the $q_\ell^\star$ term and perturbation by Δ in ℓ'', we roughly have

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \overset{\text{heuristically}}{=} O(1) \cdot \sup_{y\in\mathcal{Y},|t|\le RB} |\ell'(t, y)| \min\left\{ RB\sqrt{\gamma}, \frac{|\ell'(t, y)|}{\ell''(t, y)} \frac{1}{\sqrt{n}} \right\}, \tag{7}$$

which makes clearer the various relationships. When the ratio of $\ell'(t, y)$ to $\ell''(t, y)$ is large, estimation and optimization are intuitively hard: there is little curvature to help identify optimal parameters, while small changes in the parameter induce large changes in the loss (as $\ell'(t, y)$ is large relative to ℓ''). The allowable misspecification of the model—via the parameter γ—means that in the lower bound, an

adversary may essentially put positive mass on those points for which the ratio ℓ'/ℓ'' is large, so that one must pay this worst-case cost.

We then have the following theorem, whose proof we provide in Appendix A.

Theorem 1. *Let the loss L and family $\{p_\theta\}$ satisfy Eq. (4), where $\mathcal{X} = \{x : \|x\| \le R\}$. Consider $\Gamma = \{p_\theta\}_{\theta\in\Theta}$, where $\Theta = \{\theta : \|\theta\| \le B\}$. Then*

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \ge \frac{1}{4\sqrt{n}}\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma).$$

Using the heuristic display (7) above can provide some intuition. When $|\ell'(t,y)|/\ell''(t,y) \gtrsim \sqrt{n}$, so that the problem has little curvature, the (heuristic) linearity constant (7) scales as $\sup_{t,y}|\ell'(t,y)|RB$, which gives the lower bound $\sup_{t,y}\frac{|\ell'(t,y)|RB}{\sqrt{n}}$; this is the familiar worst-case minimax bound for stochastic convex optimization with Lipschitz objective on a compact domain [1]. As the worst-case constructions look very little like standard prediction problems, one might hope (at least in the absence of misspecification) to achieve better rates; Theorem 1 helps to delineate problems where this may be impossible.

2.2 Examples with the logarithmic loss

It is instructive to consider a few examples to build intuition for Theorem 1 beyond the heuristic (7), as the linearity constant as defined may be somewhat challenging to work with. As we shall see, however, its generality allows exploration of many losses, including various scoring rules [16]; for this section, we focus on the common logarithmic loss for four well-known exponential family models. In what follows we use the following notation: For a set Ω such that $f, g : \Omega \to \mathbb{R}$ we write $f \gtrsim g$ if there exists a finite numerical constant C such that for any $\omega \in \Omega$, $f(\omega) \ge Cg(\omega)$ and we write $f \asymp g$ if $f \gtrsim g \gtrsim f$.

For our first example, we consider logistic regression, showing that if we must play proper predictions p_θ, parametric $1/n$ rates are impossible until n is very large or if the radii R and B are small.

Example 1 (Logistic regression)**:** For logistic regression with logarithmic loss, we have $\mathcal{Y} = \{-1, 1\}$, $p_\theta(y \mid x) = \frac{1}{1+\exp(-y\theta^T x)}$, and $\ell(t,y) = \log(1 + e^{-ty})$, so that

$$\ell'(t,y) = \frac{-y}{1+e^{ty}} \quad \text{and} \quad \ell''(t,y) = \frac{e^{ty}}{(1+e^{ty})^2}.$$

Without loss of generality, let $y = 1$. If $RB \le 1$, then by taking $t = \frac{1}{2}BR$ and $\delta = \frac{1}{2}BR$, it is immediate that $q^\star_\ell(t,y) \gtrsim 1$, and each of $\ell'(t,y)$ and $\ell''(t,y)$ are numerical constants. Then we obtain the lower bound

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \ge c \min\left\{BR\sqrt{\gamma}, \frac{1}{\sqrt{n}}\right\},$$

so that Theorem 1 yields minimax lower bound $\min\{\frac{RB\sqrt{\gamma}}{\sqrt{n}}, \frac{1}{n}\}$.

The more interesting regime is when $RB \gg 1$—for example, in the natural case that the data and parameter radii scale with the dimension of the problem—so let us assume $RB \ge 1$. Here, take $y = -1$ and $y_0 = 1$, so that for any $\alpha \in [0,1]$ and $t \in \mathbb{R}$ we have $\mathrm{sign}(\ell'(t,y)) = 1 \ne -1 = \mathrm{sign}(\ell'(\alpha t, y_0))$. Let $\epsilon \in [0,1]$ to be chosen and set $t^2 = (1-\epsilon)\frac{R^2B^2}{2}$ (where $t \ge 0$). Then by taking $\alpha = \frac{1}{RB}$, in the definition (5) we have

$$q^\star_\ell(t,y) \ge \frac{\alpha\frac{1}{1+e^{t\alpha}}}{\alpha\frac{1}{1+e^{t\alpha}} + \frac{1}{1+e^{-t}}} = \frac{1}{1 + RB\frac{1+e^{t\alpha}}{1+e^{-t}}} \ge \frac{1}{1+(e+1)RB} \gtrsim \frac{1}{RB}$$

and $\ell'(t,y) = \frac{1}{1+e^{-t}} \ge \frac{1}{2}$. Thus for all $\delta \in [0, RB\sqrt{\epsilon/2}]$, the linearity constant has lower bound

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \ge c \min\left\{\delta\sqrt{\gamma/RB}, \frac{1}{\sup_{|\Delta|\le\delta} e^{-t+\Delta}}\frac{1}{RB\sqrt{n}}\right\}$$

where $c > 0$ is a numerical constant. Taking $\delta = RB\sqrt{\epsilon/2}$ and $\epsilon = 1/9$ then gives

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c \min\left\{\sqrt{\gamma RB}, \exp\left(3RB/(5\sqrt{2})\right) \frac{1}{RB\sqrt{n}}\right\}.$$

In particular, if $n \leq \frac{e^{6RB/5\sqrt{2}}}{R^2B^2}$, then $\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c\sqrt{\gamma RB}$, and otherwise (as $e^x/x \gtrsim e^{.99x}$ for all $x \geq 1$) $\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq \exp(RB/4)/\sqrt{n}$, giving us the minimax lower bound

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \geq c \min\left\{\frac{\sqrt{\gamma RB}}{\sqrt{n}}, \frac{\exp(2RB/5)}{n}\right\}. \tag{8}$$

We may contrast this lower bound with previous results. In the regime where $\gamma = 1$, Hazan et al. [24] show that for $R = 1$ and numerical constants $c_0, c_1 > 0$, any algorithm playing parametric predictors p_θ necessarily suffers minimax risk $\Omega(\sqrt{B/n})$ whenever $n \leq c_0 \exp(c_1 B)$. The result (8) recovers this lower bound while applying whenever $\gamma > 0$. $\Diamond$

To show some of the generality of our approach, we consider other exponential family models. The first is similar to the previous example.

Example 2 (Geometric distributions)**:** We say $Y \sim \mathsf{Geo}(\lambda)$ for some $\lambda \in (0, 1)$ if Y has support $\{0, 1, 2, \ldots, \}$ and $P(Y = y) = \lambda(1 - \lambda)^y$. We model this via $Y \mid x \sim \mathsf{Geo}(e^{\theta^T x}/(1 + e^{\theta^T x}))$, giving losses

$$L_{\log}(p_\theta(\cdot \mid x), y) = (y + 1) \log(1 + \exp(\theta^T x)) - \theta^T x \quad \text{and} \quad \ell(t, y) = (y + 1) \log(1 + e^t) - t.$$

We perform a quick sketch, letting $b = RB$ for shorthand, assuming that $b \geq 1$ and that $\mathrm{diam}(\mathcal{Y}) := \max\{y \in \mathcal{Y}\}$ is finite and at least 3.

First, we construct a lower bound on $q^\star_\ell(t, y)$: take $y = \max\{y \in \mathcal{Y}\}$ to be the maximum element of $\mathcal{Y}$, and set $y_0 = y$ and $t = -b$. Then setting $\alpha = -1/b$ in the definition (5) we obtain

$$q^\star_\ell(t, y) \geq \frac{-\frac{y+1}{b}\frac{1}{1+e} + \frac{1}{b}}{-\frac{y+1}{b}\frac{1}{1+e} + \frac{1}{b} - (y + 1)\frac{e^b}{1+e^b} + 1} = \frac{\frac{y+1}{1+e} - 1}{\frac{y+1}{1+e} - b + 1 + b(y + 1)\frac{e^b}{1+e^b}} \gtrsim \frac{1}{b}$$

as $b \geq 1$ and $y \geq 3$. Additionally, we have $|\ell'(t, y)| \gtrsim y$ and $\ell''(t, y) \lesssim ye^{-b}$, and so, as in the derivation in Example 1 and by setting $\delta \gtrsim b$, we obtain that there exist numerical constants $c_0, c_1 > 0$ such that

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c_0 y \min\left\{\sqrt{\gamma b}, \frac{e^{c_1 b}}{\sqrt{n}}\right\}.$$

Substituting, we obtain the analogue of inequality (8), that is,

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \geq c_0 \mathrm{diam}(\mathcal{Y}) \min\left\{\frac{\sqrt{\gamma RB}}{\sqrt{n}}, \frac{\exp(c_1 RB)}{n}\right\}.$$

Again, we see that until $n \gtrsim \exp(cRB)$, any method playing the models p_θ for points $\theta \in \Theta$ necessarily cannot converge faster than $\mathrm{diam}(\mathcal{Y})/\sqrt{n}$. $\Diamond$

Poisson regression yields a similar lower bound:

Example 3 (Poisson regression)**:** In the poisson regression problem, we model $y \in \mathbb{N}$ as $\mathsf{Poisson}(e^{\theta^T x})$, so that

$$-\log p_\theta(y \mid x) = e^{\theta^T x} - y\theta^T x + \log(y!),$$

and we may consider the loss $\ell(t, y) = e^t - yt$. We claim that in the setting of Theorem 1, where we set $\mathrm{diam}(\mathcal{Y}) = \max\{y \in \mathcal{Y}\} \geq 3$ as in Example 2, we have

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \geq c_0 \min\left\{\mathrm{diam}(\mathcal{Y})\frac{\sqrt{\gamma RB}}{\sqrt{n}}, \frac{e^{c_1 RB}\mathrm{diam}(\mathcal{Y})^2}{n}\right\}. \tag{9}$$

To see the lower bound (9), it is sufficient to lower bound the linearity constant, and we may assume that $RB \geq 2$. Our first step is to lower bound the quantity $q_\ell^\star$. Let $b = RB/2$ for shorthand, and set $t = b$, $y = \diam(\mathcal{Y})$, and $y_0 = y$, and $\alpha = -\frac{1}{b} \geq -1$, so that $\ell'(t, y) = e^t - y = e^{-b} - y < 0$ while $\alpha\ell'(\alpha t, y) = -\frac{1}{b}(e - y) > 0$. As a consequence, completely parallel to our derivation in Example 2, we obtain $q_\ell^\star(t, y) \gtrsim \frac{1}{b}$. Moreover, the bounds $\ell''(t, y) = e^t$ and the choice $\delta = b/2$ in the definition (6) yield that for numerical constants $c_0, c_1 > 0$ we have

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c_0 y \min\left\{\sqrt{\gamma b}, \frac{e^{c_1 b} y}{\sqrt{n}}\right\}.$$

Substituting this into Theorem 1 then implies the claim (9). $\Diamond$

In contrast, for linear regression problems, there is limited worst-case behavior, which is natural: the problem is always strongly convex, no matter the misspecification of the model.

Example 4 (Linear regression)**:** Again we consider the log loss, but we assume that our model is that $y \mid x \sim \mathsf{N}(\theta^T x, 1)$, so that the loss becomes $L_{\log}(p_\theta(\cdot \mid x), y) = \frac{1}{2}(\theta^T x - y)^2$ and $\ell(t, y) = \frac{1}{2}(t - y)^2$. In this case, we may take $q_\ell^\star \gtrsim 1$ in Eq. (5), and the linearity constant (6) becomes

$$\begin{aligned}\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) &\asymp \sup_{y \in \mathcal{Y}} \sup_{t^2 + \delta^2 \leq \frac{R^2 B^2}{2}} \left\{(t - y) \min\left\{\delta\sqrt{\gamma}, \frac{t - y}{\sqrt{n}}\right\}\right\} \\ &\asymp \min\left\{RB \max\{RB, \diam(\mathcal{Y})\}\sqrt{\gamma}, \frac{\max\{R^2B^2, \diam(\mathcal{Y})^2\}}{\sqrt{n}}\right\},\end{aligned}$$

yielding minimax lower bound

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \geq c \min\left\{\frac{\max\{R^2B^2, RB\diam(\mathcal{Y})\}\sqrt{\gamma}}{\sqrt{n}}, \frac{\max\{R^2B^2, \diam(\mathcal{Y})^2\}}{n}\right\}.$$

This is sharp: the stochastic gradient method achieves the bound [22], as ℓ is strongly convex and has Lipschitz constant $\max\{BR, \diam(\mathcal{Y})\}$.

With that said, this behavior—which depends on $\diam(\mathcal{Y})$—is worse than what one achieves in well-specified or stochastic settings, where the stochasticity means that rates of $1/n$ are achievable. $\Diamond$

2.3 Scoring rules and general losses

In probabilistic prediction and forecasting, one more generally may consider *scoring rules* [16], which are losses designed to engender various behaviors: honesty in eliciting predictions, calibration of forecasts, robustness, or other reasons. Typically, these induced losses are exp-concave (as we discuss in the next section, which will allow us to describe an efficient algorithm for them). For example, to achieve robustness [31, 25] (there are no unbounded losses) one might consider the squared error or Hellinger-type losses

$$L_{\rm sq}(p(\cdot \mid x), y) := \frac{1}{2}(p(y \mid x) - 1)^2 \quad \text{and} \quad L_{\rm hel}(p(\cdot \mid x), y) = (\sqrt{p(y \mid x)} - 1)^2, \tag{10}$$

neither of which is proper (so that the true distribution may not minimize the loss). Alternatively, proper scoring rules [16] are minimized by the true predictive distribution, and include the logarithmic loss and the quadratic scoring rule with loss

$$L_{\rm quad}(p(\cdot \mid x), y) := \frac{1}{2}\sum_{k \in \mathcal{Y}}(p(k \mid x) - \mathbf{1}\{k = y\})^2 = \frac{1}{2}(p(y \mid x) - 1)^2 + \frac{1}{2}\sum_{k \neq y} p(k \mid x)^2. \tag{11}$$

As these scoring rules are differentiable and at least $\mathcal{C}^2$ on $p \in (0, 1)$, an argument by the delta method [36] shows that in well-specified cases, one expects to achieve convergence at rate $1/n$. Moreover, as we will

see in the next section, aggregating algorithms can achieve regret scaling as $\log n$ for each of these loss measures.

Yet, at least in an asymptotic sense, the minimax bound in Theorem 1 shows that this is unachievable with misspecification. Here, because of the complexity of the losses and resulting calculations, we take a completely asymptotic perspective, saying that we have an *asymptotic rate* $\mathsf{r}(n)$ *minimax lower bound* if $\mathsf{r}(n) \to 0$ as $n \to \infty$, while

$$\liminf_{n\to\infty} \frac{\mathfrak{M}_n(\Theta, \Gamma, \gamma)}{\mathsf{r}(n)} > 0$$

for all $\gamma > 0$. Each of the bounds in Examples 1–4 is then asymptotic rate $\mathsf{r}(n) = \frac{1}{n}$. Once we move beyond the log loss to alternative scoring rules, however, the rate $\mathsf{r}(n) = \frac{1}{n}$ is no longer achievable with misspecification if $\Gamma = \{p_\theta\}$ are the proper models.

As usual we consider losses taking the form $L(p_\theta(\cdot \mid x), y) = \ell(\theta^T x, y)$ for some scalar induced loss $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. We restrict our focus to losses for which no *universally perfect* prediction exists, meaning that if $t \in \mathbb{R}$ and $y \in \mathcal{Y}$ satisfy $\ell'(t, y) \neq 0$, there exists $y_0 \in \mathcal{Y}$ such that $\ell'(t, y)\ell'(t, y_0) < 0$. We have the following result, whose proof we provide in Appendix A.4.

Proposition 2. *Assume that the scalar loss allows no universally perfect prediction. If there exists $|t| \le RB/2$ and $y \in \mathcal{Y}$ such that $\ell'(t, y) \neq 0$ while $\ell''(t, y) = 0$ and $\ell(\cdot, y)$ is $\mathcal{C}^3$ near t, then the prediction family $\Gamma = \{p_\theta\}_{\theta\in\Theta}$ has asymptotic rate $\mathsf{r}(n) = n^{-3/4}$ minimax lower bound.*

Roughly, the result in the proposition is simple: if the induced scalar loss ℓ is not convex in t, then proper predictions cannot be rate-optimal (as rates scaling as $1/n$ are achievable here).

While it is possible in some cases to achieve explicit constants—for example, for logistic regression this is relatively straightforward—in general it is somewhat tedious. Nonetheless, we have the following result, whose tediousness in verification precludes our including a formal proof, but essentially we need simply note that the induced losses $\ell(t, y)$ for each problem are non-convex in t but are smooth.

Corollary 3. *Let $\Gamma = \{P_\theta\}$ be any of the logistic-, geometric-, poisson-, or linear-regression families of predictive densities. Then for any of the squared L_{sq}, Hellinger L_{hel}, or quadratic L_{quad} losses (Eqs. (10) and (11)), $\mathfrak{M}_n(\Theta, \Gamma, \gamma)$ has asymptotic rate lower bound $\mathsf{r}(n) = \frac{1}{n^{3/4}}$ for any $\gamma > 0$.*

Theorem 1 and its consequences via Proposition 2 assert that *any* algorithm returning elements of the parametric family $\{p_\theta\}_{\theta\in\Theta}$ must suffer when misspecification is possible. These results are information-theoretic, and as such, we see that in situations where misspecification is possible, to perform better it is essential that the family Γ of allowable distributions be improper.

3 Robustness via Improper Learning

While proper algorithms evidently must suffer losses when they are misspecified, we now show that simply by considering $\Gamma = \text{Conv}\{p_\theta\}_{\theta\in\Theta}$ we can sidestep the lower bound of Theorem 1. Specifically, we consider Vovk's Aggregating Algorithm and show that this provides stability to misspecification. We present the algorithm in an online setting, though standard online-to-batch conversion techniques [5] extend the result to the stochastic optimization setting in which we have proved each of our lower bounds.

To give a regret bound, we continue our usual focus by considering losses L and families $\{p_\theta\}_{\theta\in\Theta}$ for which we can write $L(p_\theta(\cdot \mid x), y) = \ell(\theta^T x, y)$. We restrict ourselves somewhat to considering *mixable* losses, where for some $\eta > 0$, the function $p \mapsto \exp(-\eta L(p, y))$ is concave over the collection $\mathcal{P}(\mathcal{Y})$ of distributions on $\mathcal{Y}$. This constant η guaranteeing the exp-concavity of L bounds the *mixability* constant, which allows one to obtain "fast rates" via exponentially-weighted averaging in many online learning problems [40, 5, 37]. In Table 1, we record the mixability constants for several example losses—each of

Algorithm 1 Vovk's Aggregating Algorithm (Online Setting)

Define $\Gamma = \mathrm{Conv}\{p_\theta\}_{\theta\in\Theta}$ and let η be some fixed value > 0.
For $t = 0, \dots, n$
  Nature reveals x_t.
  Define $d\widehat{\mu}^{\mathsf{Vovk}}_{t,\eta}(\theta) \propto \exp(-\eta \sum_{s=0}^{t-1} L(P_\theta(y_s \mid x_s)))$.
  Decision Maker plays $\widehat{P}^{\mathsf{Vovk}}_{t,\eta} := \int_{\theta\in\Theta} P_\theta(\cdot \mid x_t) d\widehat{\mu}^{\mathsf{Vovk}}_{t,\eta}(\theta)$.
  Nature reveals y_t and Decision Maker suffers loss $L(\widehat{P}^{\mathsf{Vovk}}_{t,\eta}(y_t \mid x_t))$.

	$L_{\log}$	L_{sq}	L_{hel}	L_{quad}
Mixability Constant η	1	1	3	1/2

Table 1. Mixability constants for common losses in Eqs. (11) and Eqs. (10), where $L_{\log}(p, y) = -\log p(y)$, $L_{\mathrm{sq}}(p, y) = \frac{1}{2}(p(y) - 1)^2$, $L_{\mathrm{hel}}(p, y) = (\sqrt{p(y)} - 1)^2$, and $L_{\mathrm{quad}}(p, y) = \frac{1}{2}\|p(\cdot) - e_y\|_2^2$. See Appendix B.1.

which we touch on in Sec. 2.3—showing that Vovk's aggregating algorithm achieves logarithmic regret for any of them once we apply the coming convergence result.

Corollary 4 (Foster et al. [14], Theorem 1)**.** *Let $\widehat{P}^{\mathsf{Vovk}}_{t,\eta}$ be as defined above. Let $\mathrm{Lip}_\ell(T)$ be the Lipschitz constant of ℓ restricted to $[-T, T] \times \mathcal{Y}$. Then for any sequence $(x_i, y_i)_{i=1}^n$ and $\theta^\star \in \Theta$,*

$$\mathsf{Reg}_n(\widehat{P}^{\mathsf{Vovk}}_{n,\eta}, p_{\theta^\star}) \le 5\frac{d}{\eta} \log\left(\frac{\mathrm{Lip}_\ell(RB)n}{d} + e\right).$$

Using Corollary 4 and Theorem 1, we see that the aggregating algorithm typically provides a stronger convergence guarantee than algorithms constrained to $\{p_\theta\}_{\theta\in\Theta}$ can attain. In each of Examples 1–3, the lower bounds necessarily suffer exponential dependence $\exp(\Omega(1)RB)$ as $n \to \infty$, and so as long as the Lipschitz constant $\mathrm{Lip}_\ell(RB)$ is not super-exponential in RB, Corollary 4 guarantees better convergence. Even more, in some cases—for example, when using the general (potentially non-convex in θ) losses as in Sec. 2.3—even for fixed radii R, B we have $\sqrt{n}\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \to \infty$ as $n \to \infty$. In this case, Corollary 4 even guarantees a better asymptotic rate in n.

3.1 Lower bounds for arbitrary improper algorithms

An important question is whether the Aggregating Algorithm 1 achieves optimal rates when the misspecification parameter γ changes. As the coming Theorem 5 shows, Corollary 4 is tight to within logarithmic factors, as we can show an (asymptotic) lower bound of d/n even for well-specified families of generalized linear models. The theorem also shows that while playing in the convex hull of a parameterized family $\{p_\theta\}_{\theta\in\Theta}$ allows more powerful mechanisms, in natural (well-specified) scenarios, this extra power is no panacea: the upper bound on the risk of the Aggregating Algorithm 1 is tight to a logarithmic factor over all algorithms that may play arbitrary elements in the convex hull (or even algorithms playing any probability distribution). Indeed, let Γ_{all} be the collection of *all* probability distributions on $Y \mid X$, so that an algorithm may play an arbitrary distribution (which is of course larger than $\mathrm{Conv}\{p_\theta\}_{\theta\in\Theta}$). We then take the risk to be the expected logarithmic loss, where for a conditional distribution $p(y \mid x)$ we define

$$\mathsf{Risk}_P(p) := \mathbb{E}_P[L_{\log}(p(\cdot \mid X), Y)] = -\mathbb{E}_P[\log p(Y \mid X)],$$

and we let $\mathsf{Risk}^\star_P = \inf_p \mathsf{Risk}_P(p)$ be the smallest risk across all predictive distributions $p(y \mid x)$. When the models p_θ are not misspecified, i.e., $\gamma = 0$, we have $\mathsf{Risk}^\star_{P_\theta} = \mathsf{Risk}_{P_\theta}(p_\theta)$, and we have the following lower bound.

Theorem 5. *Let $\{p_\theta\}_{\theta\in\Theta}$ be a generalized linear model of the form*

$$dP_\theta(y \mid x) = \exp(y\theta^T x - A(\theta^T x))d\nu(y),$$

where ν is a base measure and $X \sim \mathsf{Uni}(\{-1,1\}^d)$, 0 is in the interior of $\operatorname{dom} A$, and let Θ contain $\mathbf{0}$ in its interior. Then for the log loss $L_{\log}$, there exists a numerical constant $c > 0$ such that for all large enough n

$$\mathfrak{M}_n(\Theta, \Gamma_{\text{all}}, 0) = \inf_{\widehat{p}_n} \sup_{\theta\in\Theta} \mathbb{E}^n_{P_\theta}\left[\mathsf{Risk}_{P_\theta}(\widehat{p}_n) - \mathsf{Risk}^\star_{P_\theta}\right] \ge c\frac{d}{n}.$$

See Appendix B.2 for the proof of Theorem 5. As we essentially only care about the conditional distribution $p_\theta(y \mid x)$, here we chose the marginal over X to be uniform for convenience; other choices suffice as well.

Comparing the lower bound Theorem 5 provides with the regret bound in Corollary 4, we see that holding RB constant (and $\mathrm{Lip}_\ell(RB)$ constant) then for risk functional $\mathsf{Risk}_P(p) := \mathbb{E}_P[L_{\log}(p(\cdot \mid X), Y)]$, a standard online-to-batch conversion (or Jensen's inequality) implies

$$\mathbb{E}\left[\mathsf{Risk}_P(\widehat{p}_n^{\mathsf{Vovk}}) - \inf_{\theta^\star\in\Theta} \mathsf{Risk}_P(p^\star_\theta)\right] \le (5 + o(1))\frac{d\log n}{n}$$

as $n \to \infty$, where $o(1) \to 0$ hides problem-dependent constants. To within a factor $O(1)\log n$, then, Vovk's aggregating algorithm is generally unimprovable.

3.2 Asymptotically aggregating to point predictions

While in general the ability to play elements in $\mathrm{Conv}\{p_\theta\}_{\theta\in\Theta}$ could (in principle) yield much better performance than any individual element p_θ for $\theta \in \Theta$, in a sense, the aggregating algorithm is only performing a small amount of averaging to substantially increase its robustness. Indeed, when the risk minimization problem at hand is classical—the loss has continuous derivatives and the population risk Risk_P is strongly convex in a neighborhood of its minimizer $\theta^\star$—then we can show that Vovk's aggregating algorithm asymptotically plays points very close to $\{p_\theta\}_{\theta\in\Theta}$. That is, in "nice" cases, the aggregating algorithm more or less behaves as the empirical risk minimizer, which is asymptotically optimal [13]. In these cases, stochastic gradient methods (which are necessarily proper, as they optimizer over θ) similarly achieve optimal asymptotic rates [13], and sometimes similarly strong finite sample rates [3].

We make this more formal via a generalized Bernstein von-Mises Theorem, which shows that when a unique minimizer exists, the density $d\widehat{\mu}^{\mathsf{Vovk}}_{n,\eta}$ in Alg. 1 converges to a normal density centered at the empirical minimizer $\widehat{\theta}_n \in \Theta$ with covariance operator shrinking at the rate $1/n$. Such posterior limiting normality results are relatively well-known: see [36, 28]. In our setting, rather than considering just the posterior distribution—as our models may be misspecified, so that a posterior is less sensible—we consider distributions over Θ of the form of $\widehat{\mu}^{\mathsf{Vovk}}_{n,\eta}$; when L is the log-loss, this is the usual posterior. We first define the class of families and losses for which Theorem 6 holds.

Definition 3.1. *The family and loss pair $(\{p_\theta\}_{\theta\in\Theta}, L)$ is* Bernstein von-Mises generalizable *if there exist $\epsilon_1, \epsilon_2 > 0$ such that the risk $\mathsf{Risk}_P(\theta) := \mathbb{E}_P[L(P_\theta(Y \mid X))]$ satisfies the following conditions:*

(i) The minimizer $\theta^\star = \operatorname{argmin}_{\theta\in\Theta} \mathsf{Risk}_P(\theta)$ is unique and has positive definite Hessian $\nabla^2\mathsf{Risk}_P(\theta^\star) \succ 0$.

(ii) On the ϵ_1-ball around $\theta^\star$, $\theta^\star + \epsilon_1\mathbb{B}_2^d$, the loss $\theta \mapsto L(P_\theta(\cdot \mid x), y)$ is $M_{\mathrm{Lip},0}(x,y)$ Lipschitz and has $M_{\mathrm{Lip},2}(x,y)$-Lipschitz Hessian, where $\mathbb{E}[M_{\mathrm{Lip},0}(X,Y)^2] < \infty$ and $\mathbb{E}[M_{\mathrm{Lip},2}(X,Y)] < \infty$.

(iii) For all $\theta \in \Theta \setminus \{\theta^\star + \epsilon_1\mathbb{B}_2^d\}$, we have $\mathsf{Risk}_P(\theta) \ge \mathsf{Risk}_P(\theta^\star) + \epsilon_2$.

When $\theta \mapsto L(P_\theta(\cdot \mid x), y)$ is convex, condition (iii) is redundant given the others, and the other conditions of Definition 3.1 hold for generalized linear models. Under the conditions Definition 3.1 specifies, we then obtain the following convergence guarantee, whose proof we provide in Appendix B.3.

Theorem 6 (Generalized Bernstein von-Mises)**.** *Let the pair $(\{p_\theta\}_{\theta\in\Theta}, L)$ be Bernstein von-Mises generalizable (Definition 3.1), and for $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ define $\mathsf{Risk}_n(\theta) = \frac{1}{n}\sum_{i=1}^n L(P_\theta(\cdot \mid X_i), Y_i)$. Assume Θ is compact and $\theta^\star \in \operatorname{int}\Theta$. Let $\widehat{\theta}_n := \operatorname{argmin}_{\theta\in\Theta} \mathsf{Risk}_n(\theta)$ and $\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}}$ be as defined in Vovk's Aggregating Algorithm. Then*

$$\left\|\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - \mathsf{N}\left(\widehat{\theta}_n, \frac{1}{n}\nabla^2 \mathsf{Risk}_n(\widehat{\theta}_n)^{-1}\right)\right\|_{\text{TV}} \xrightarrow[P]{a.s.} 0.$$

Using the theorem and its proof, we can also (under a minor continuity condition) establish a convergence guarantee showing roughly that the aggregating algorithm asymptotically plays essentially the empirical point estimator. We consider the following assumption.

Assumption 1. *There exists a neighborhood B of $\theta^\star$ such that the log-likelihood $\theta \mapsto \log p_\theta(y \mid x)$ is $\operatorname{Lip}_p(x,y)$-Lipschitz on B, and $\operatorname{Lip}_p(x) := \sup_{\theta\in B}\int_{\mathcal{Y}} \operatorname{Lip}_p(x,y) dP_\theta(y \mid x) < \infty$ for each x.*

We then have the following corollary, whose proof we provide in Appendix B.4.

Corollary 7. *In addition to the conditions of Theorem 6, let Assumption 1 hold. Then for each $x \in \mathcal{X}$,*

$$\left\|\widehat{P}_{n,\eta}^{\mathsf{Vovk}}(\cdot \mid x) - P_{\widehat{\theta}_n}(\cdot \mid x)\right\|_{\text{TV}} \overset{a.s.}{\to} 0.$$

Roughly, Theorem 6 and Corollary 7 show that the aggregating algorithm 1 is asymptotically constrained to making predictions in $\{p_\theta\}_{\theta\in\Theta}$, at least in non-adversarial cases. In a sense, then, the aggregating algorithm 1 is not taking full advantage of its improperness: while it can return any distribution in $\operatorname{Conv}\{p_\theta\}_{\theta\in\Theta}$, it (eventually) is nearly playing elements of $\{p_\theta\}_{\theta\in\Theta}$. While this is optimal in some cases (Theorem 5), the question of how to efficiently and optimally return predictions in $\operatorname{Conv}\{p_\theta\}_{\theta\in\Theta}$ remains open and a natural direction for future work.

4 Experiments and Implementation Details

Before discussing our experiments, we make a few remarks on the computability of Vovk's Aggregating Algorithm. Whenever $L(P_\theta(\cdot \mid x), y)$ is convex in θ and β-smooth, there exists an algorithm [14] approximating $\widehat{P}_{n,\eta}^{\mathsf{Vovk}}$ that achieves the regret bound in Corollary 4 to within an additive factor $1/n$, and the algorithm is polynomial in $(RB, d, \operatorname{Lip}_\ell(RB), n)$. Yet these algorithms are still computationally intensive; assuming our theoretical results are predictive of actual performance, one might expect that aggregating-type strategies could still yield improvements over standard empirical risk minimization. Indeed, Jézéquel et al. [27] take the computational difficulty of the approximating algorithms in the paper [14] as motivation to develop an efficient improper learning algorithm for the special case of logistic regression, which (roughly) hedges its predictions by pretending to receive both positive and negative examples in future time steps, constructing a loss that depends explicitly on the new data x_t; Jézéquel et al. show that it achieves a regret bound with a multiplicative RB factor of the logarithmic regret in Corollary 4. It is unclear how to extend this approach to situations in which the cardinality $|\mathcal{Y}|$ of $\mathcal{Y}$ is much larger than 1, though this is an interesting question for future work. In our experiments, we take a heuristic approach, focusing on the risk minimization setting, and perform aggregation of subsampled maximum likelihood estimators; this approach is reminiscent of the subsampled and bootstrapped estimators [43, 29], but we use aggregation as in Alg. 1 to weight predictions. We call the procedure AHA (**A** **H**euristic **A**ggregation) for short.

Algorithm 2 AHA (**A** **H**euristic **A**ggregation)

Input: $\{(x_i, y_i)\}_{i=1}^n$ and parameter radius B
Output: $P_{n,B}^{\mathsf{mix}}$

For $k = 1, \cdots, K$

$S_k \leftarrow$ random subset of the data of size $|S_k| = 2n/3$

$\widehat{\theta}_n^k \leftarrow \operatorname{argmin}_{\|\theta\| \le B} \sum_{(x,y) \in S_k} L(p_\theta(y \mid x))$

$\mu_n^k \leftarrow \exp(\sum_{i=1}^n -L(p_{\widehat{\theta}_n^k}(y_i \mid x_i)))$

$P_{n,B}^{\mathsf{mix}} \leftarrow \big(\sum_{k=1}^K \mu_n^k p_{\widehat{\theta}_n^k}\big) / \big(\sum_{k=1}^K \mu_n^k\big)$

Our results suggest that when performing a probabilistic forecasting task with parameterized model $\{p_\theta\}_{\theta \in \Theta}$, returning the mixture distribution from Vovk's Aggregating Algorithm should be more robust to misspecification than an algorithm which returns $P_{\widehat{\theta}}$; we thus expect Algorithm 2 should exhibit more robustness as well. To that end, we consider two experiments: the first a synthetic experiment with linear regression, where we may explicitly control the degree of misspecification, and the second a logistic regression problem on real digit recognition data, where we mix two populations and we expect (roughly) that a model should do well on one, but may be missing important aspects of the other.

Improper Linear Regression For the synthetic data, we let $X \in \mathbb{R}^d$ be an observed covariate and $H \in \mathbb{R}$ a hidden variable, and for $\tau \in \mathbb{R}_+$ we let y have density

$$p_\tau(y \mid X = x, H = h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y - (x^T\theta^\star + \tau h)^2}.$$

We fix the dimension $d = 10$ and let $\theta^\star \in \mathbb{R}^d$ be uniform on $\mathbb{S}^{d-1}$; we generate data by drawing $(X, H) \sim \mathsf{N}(0, I_d) \times \mathsf{N}(0, 1)$. We then use the parametric model $\{p_\theta\}_{\theta \in \Theta}$ to model $Y \mid X = x \sim \mathsf{N}(\theta^T x, 1)$, which is misspecified when $\tau > 0$. As τ grows larger—increasing misspecification—we expect greater differences between the M.L.E. $\widehat{\theta}_n$ and the AHA Algorithm 2.

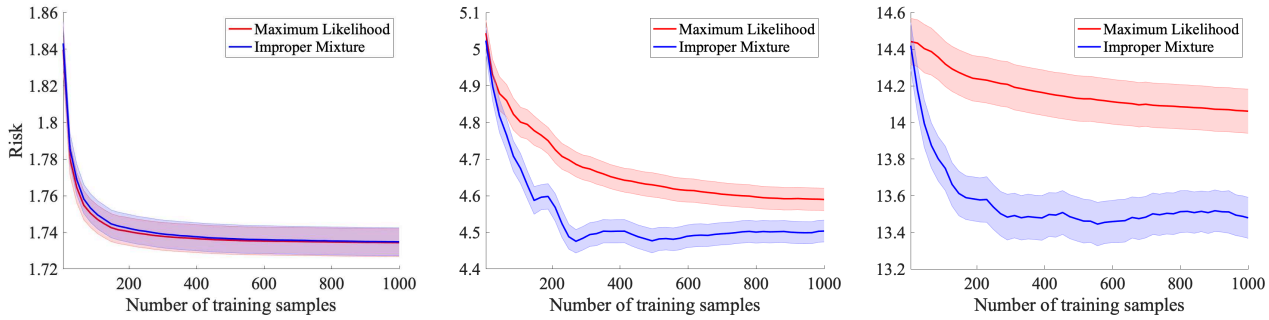
Logistic Regression We consider the MNIST handwritten digits [30], where we mix in typed digits; as our base featurization, we use a standard 7-layer convolutional neural network trained on the MNIST data, so that the typewritten digits (roughly) represent a misspecified sub-population, and as the proportion τ of typewritten digits increases, we expect increasing misspecification. We consider a simplified binary version of this problem, where we seek to distinguish digits 3 and 8, and we use a logistic regression model $p_\theta(y \mid x) = (1 + e^{-yx^T\theta})^{-1}$ with log loss.

Experiment For both linear regression and logistic regression, we conduct the following experiment: For a training sample size n and parameter radius B, we compute the constrained MLE

$$\widehat{\theta}_n := \operatorname*{argmin}_{\|\theta\| \le B} \sum_{i=1}^n L(P_\theta(\cdot \mid x_i), y_i)$$

and return $P_{\widehat{\theta}_n}$, and also compute $P_{n,B}^{\mathsf{mix}}$ as the output of Alg. 2 with resampling size $K = 20$ for the improper linear regression experiment and $K = 10$ for the MNIST experiment. We use a held-out test set of size $N = 5000$ to approximate the risk $\mathsf{Risk}(p)$ of the returned conditional probability p. We plot how this approximated risk decays as we increase training sample size n up to 1000 for improper linear regression and up to 200 for logistic regression.

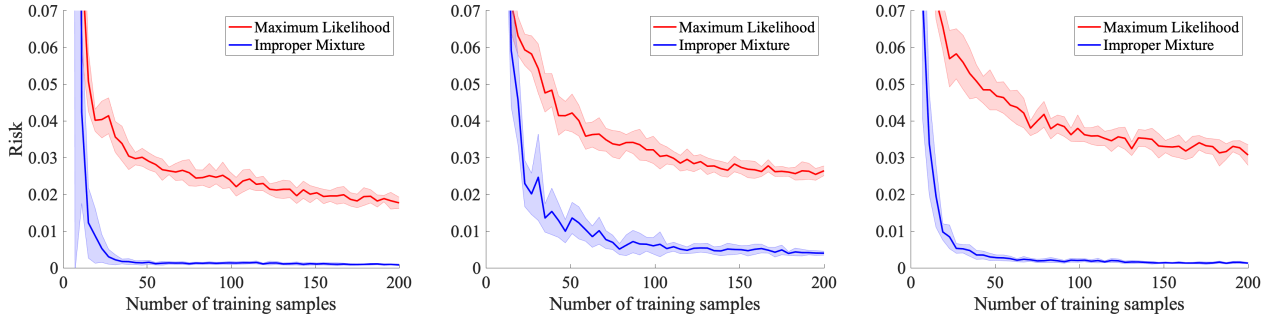
Within each experiment, we implement several regularization schedules. We test $B = c$, $B = c \log n$, $B = c\sqrt{n}$, and $B = cn$ for $c \in \{0.1, 0.2, 1\}$. In Figures 1 and 2 *we only show the results for the best choice*



(a) $\tau = 0$ (b) $\tau = 2.5$ (c) $\tau = 5$

Figure 1. Linear Regression, Synthetic Data. As misspecification τ increases, the improper learning algorithm AHA (Alg. 2) outperforms the best constrained MLE.

of B according to the performance of the Maximum Likelihood Estimator. We repeat the experiment 100 times on the synthetic data and 10 times on the real dataset and average the results. We run the experiment for $\tau = 0, 2.5, 5$ on the synthetic dataset and $\tau = 0\%, 5\%, 20\%$ for the real dataset.



(a) $\tau = 0\%$ (b) $\tau = 5\%$ (c) $\tau = 20\%$

Figure 2. Logistic Regression, MNIST Data mixed with typed data. As misspecification τ increases, the improper learning algorithm AHA's performance remains stable, while the best regularized MLE's performance worsens.

The results of Figure 1 and Figure 2 are consistent with our expectations: as the magnitude of misspecification (as measured by $\tau \geq 0$) increases, the gap in performance between the maximum likelihood estimator and the aggregated solution increases. Even more, if we may be so bold, the results suggest that using a subsampling and aggregation strategy as in Alg. 2 may be a useful primitive for other learning problems; we leave this as a possibility for future work.

5 Discussion

This work takes steps toward addressing the fundamental and practically important challenge of the cost of inaccurate modeling. While modeling assumptions are ubiquitous throughout statistics, machine learning, and data science—allowing analyses that demonstrate fast convergence rates, efficient algorithms, interpretable conclusions—most such assumptions are (at least) slightly flawed. This misspecification can have downsides: in addition to perhaps faulty conclusions from a faulty model, even convergence rates of estimators may degrade. This adds a wrinkle to data-modeling tasks: not only must we choose a model that closely fits the data, but we must be mindful of the cost of model misspecification, as this cost is not uniform across models. Our development of the linearity constant $\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma)$ in Eq. (6) of the model family gives a reasonably concise description of potential sensitivity to misspecification for

many model families.

Yet as we additionally consider, for probabilistic prediction problems aggregation strategies can at least ameliorate these challenges. Of course, aggregation approaches are familiar throughout statistical learning [35, 10, 37], but we believe their potential for improvement *beyond* "optimal" point estimators remains unexplored; our results provide one lens for viewing this problem.

References

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

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[3] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[4] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27(6): 1865–1895, 1999.

[5] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[6] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366, 2002.

[7] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203, 1999.

[8] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, Jan. 1991.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[10] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72:39–61, 2008.

[11] A. DeSantis, G. Markowsky, and M. Wegman. Learning probabilistic prediction functions. In *COLT*, pages 312–328, 1988.

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

[13] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, To Appear, 2020.

[14] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, pages 167–208, 2018.

[15] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[16] T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[17] C. W. Granger and R. Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1984.

[18] P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[19] S. G. Hall and J. Mitchell. Combining density forecasts. *International Journal of Forecasting*, 23 (1):1–13, 2007.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[21] D. Haussler and A. Barron. How well do bayes methods work for on-line prediction of +1;-1 values? In *Proceedings of the Third NEC Symposium on Computation and Cognition. SIAM*, 1992.

[22] E. Hazan and S. Kale. An optimal algorithm for stochastic strongly convex optimization. In *Proceedings of the Twenty Fourth Annual Conference on Computational Learning Theory*, 2011. URL http://arxiv.org/abs/1006.2425.

[23] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[24] E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the Twenty Seventh Annual Conference on Computational Learning Theory*, pages 197–209, 2014.

[25] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

[26] R. A. Jacobs. Methods for combining experts' probability assessments. *Neural computation*, 7(5): 867–888, 1995.

[27] R. Jézéquel, P. Gaillard, and A. Rudi. Efficient improper learning for online logistic regression. *arXiv:2003.08109*, 2020.

[28] B. Kleijn and A. Van der Vaart. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[29] A. Kleiner, A. Talwalkar, P. Sarkar, and M. Jordan. Bootstrapping big data. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[30] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks*, pages 53–60, 1995.

[31] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for non-convex losses. *Annals of Statistics*, 46(6):2747–2774, 2018.

[32] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6): 2124–2147, 1998.

[33] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.

[34] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

[35] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32 (1):135–166, 2004.

[36] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[37] T. Van Erven, P. Grunwald, N. A. Mehta, M. Reid, R. Williamson, et al. Fast rates in statistical and online learning. 2015.

[38] V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, 1990.

[39] V. G. Vovk. Universal forcasting algorithms. *Information and Computation*, (96):245–277, 1992.

[40] V. G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, Apr. 1998.

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

[42] K. Yamanishi. A loss bound model for on-line stochastic prediction algorithms. *Information and Computation*, 119(1):39–54, 1995.

[43] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

[44] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Proof of Theorem 1

Before we give the proof of the theorem proper, we first recall Le Cam's method. As we consider the excess loss, central to our development is the following separation quantity [cf. 12, Sec. 5].

Definition A.1 (Separation)**.** *Let $f_1 : \Theta \to \mathbb{R}$, $f_2 : \Theta \to \mathbb{R}$. Their* separation *with respect to Θ is*

$$\mathsf{sep}(f_1, f_2, \Theta) := \sup \left\{ \varepsilon \geq 0 \mid \begin{array}{ll} f_1(\theta) \leq \inf_{\theta \in \Theta} f_1(\theta) + \varepsilon & \textit{implies} \quad f_2(\theta) > \inf_{\theta \in \Theta} f_2(\theta) + \varepsilon \\ f_2(\theta) \leq \inf_{\theta \in \Theta} f_2(\theta) + \varepsilon & \textit{implies} \quad f_1(\theta) > \inf_{\theta \in \Theta} f_1(\theta) + \varepsilon, \end{array} \textit{ all } \theta \in \Theta \right\}.$$

This separation measures the extent to which minimizing a function f_1 means that one cannot minimize a function f_2, and by a standard reduction of estimation and optimization to testing—if one can optimize well, then one can decide whether one is optimizing f_1 or f_2—we have Le Cam's method. (See [12, Sec. 5.2] for this specific form.)

Lemma 8 (Le Cam's Method)**.** *Let $v \in \{\pm 1\}$ and P_v be arbitrary distributions on a set $\mathcal{Z}$ and $f_v : \Theta \to \mathbb{R}$ be functions similarly indexed by $v \in \{\pm 1\}$, where $f_v^\star = \inf_{\theta \in \Theta} f_v(\theta)$. Then*

$$\inf_{\widehat{\theta}} \max_{v \in \{-1,1\}} \mathbb{E}_{P_v^n} \left[f_v(\widehat{\theta}(Z_1, \ldots, Z_n)) - f_v^\star \right] \geq \mathsf{sep}(f_1, f_{-1}, \Theta) \left(1 - \sqrt{\frac{n}{2} D_{\rm kl}\left(P_1 \| P_{-1}\right)} \right),$$

where the infimum is over $\widehat{\theta} : \mathcal{Z}^n \to \Theta$ and the expectation is over $Z_i \stackrel{\rm iid}{\sim} P_v$.

To use Lemma 8 to prove lower bounds, then, the key is to show that for a given loss L, there are distributions P_1, P_{-1} that induce a large separation in the risks Risk_{P_v} while having small KL-divergence. The basic approach, familiar from other lower bounds [12, 41], is to show that for some constants $0 < c_0, c_1 < \infty$ and a power $\beta \geq 0$, we can choose $P_{\pm 1}$ to scale with a desired rate ε via

$$\mathsf{sep}(\mathsf{Risk}_{P_1}, \mathsf{Risk}_{P_{-1}}, \Theta) \geq c_0 \varepsilon^\beta \quad \text{while} \quad D_{\rm kl}\left(P_1 \| P_{-1}\right) \leq c_1 \varepsilon^2.$$

Given these separation and divergence bounds, it is then evidently the case that we may choose $\varepsilon^2 = \frac{1}{2c_1 n}$, which immediately yields a lower bound via Lemma 8 scaling as

$$c_0 \left(\frac{1}{2c_1 n} \right)^{\beta/2}.$$

Thus any lower bounds we prove become larger as the separation rate β decreases or constant c_0 grows. The next lemma does precisely this, though there is some sophistication required because of the different constraints on our losses.

Lemma 9. *Let the loss take the form $L(p_\theta(y \mid x)) = \ell(\theta^T x, y)$. Let $\varepsilon \in [0, \frac{3}{5}]$, $y \in \mathcal{Y}$, and $t \in \mathbb{R}$, and $q_\ell^\star(t, y)$ be as in definition (5). Assume t and $\delta \geq 0$ jointly satisfy*

$$\sup_{|\Delta| \leq \delta} \delta \ell''(t + \Delta, y) \leq \varepsilon q_\ell^\star(t, y) |\ell'(t, y)| \quad \textit{and} \quad 2(t^2 + \delta^2) \leq R^2 B^2. \tag{12}$$

Then for any $\mathcal{X} \supset \{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\}$, there exist distributions $\{P_{\pm 1}\}$ on $\mathcal{X} \times \mathcal{Y}$ such that

$$\mathsf{sep}(\mathsf{Risk}_{P_1}, \mathsf{Risk}_{P_{-1}}, \Theta) \geq \frac{q_\ell^\star(t, y)}{2} |\ell'(t, y)| \delta \cdot \varepsilon$$

while

$$D_{\rm kl}\left(P_1 \| P_{-1}\right) \leq q_\ell^\star(t, y) \varepsilon^2.$$

We prove Lemma 9 in Appendix A.2.

Now we leverage Lemma 9 to provide a minimax risk bound over γ variation distance perturbations. The key here is that the family $\{P_\theta\}$ restricts only conditional distributions—the marginal distribution over $X \in \mathcal{X}$ may be arbitrary—allowing us to give appropriate mixtures.

Lemma 10. *Assume that $\mathcal{X} \supset \{x \in \mathbb{R}^d \mid \|x\|_2 \le R\}$ and let $P_{\pm 1}$ be distributions on $\mathcal{X} \times \mathcal{Y}$. Let $\gamma \in [0, 1]$. Then there exists a distribution $P_0 \in \{P_\theta\}_{\theta \in \Theta}$ such that for $Q_{\pm\gamma} := (1-\gamma)P_0 + \gamma P_{\pm 1}$,*

$$\mathsf{sep}(\mathsf{Risk}_{Q_\gamma}, \mathsf{Risk}_{Q_{-\gamma}}, \Theta) = \gamma \, \mathsf{sep}(\mathsf{Risk}_{P_1}, \mathsf{Risk}_{P_{-1}}, \Theta) \quad \text{and} \quad D_{\rm kl}\left(Q_\gamma \| Q_{-\gamma}\right) \le \gamma D_{\rm kl}\left(P_1 \| P_{-1}\right).$$

See Appendix A.3 for the short proof of the result.

With Lemma 10 in hand we can now prove Theorem 1.

A.1 Proof of Theorem 1 proper

First, recalling the perturbed minimax risk from Definition 1.1,

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) := \inf_{\widehat{p}_n \in \Gamma} \sup_{\theta \in \Theta} \sup_{P : \|P - P_\theta\|_{\rm TV} \le \gamma} \mathbb{E}_{P^n}[\mathsf{Risk}_P^\Theta(\widehat{p}_n)],$$

where the infimum is over all procedures. Now, let $(\varepsilon, y, t, \delta)$ be any collection satisfying the conditions of Lemma 9 and $\{P_{\pm 1}\}$ be the distributions the lemma guarantees exist. Additionally, let $Q_{\pm\gamma}$ be the perturbed distributions Lemma 10 provides, so that there exists $P_0 \in \{P_\theta\}$ such that $\|P_0 - Q_{\pm\gamma}\|_{\rm TV} \le \gamma \|P_0 - P_{\pm 1}\|_{\rm TV} \le \gamma$. Then we immediately obtain

$$\begin{aligned}
\mathfrak{M}_n(\Theta, \Gamma, \gamma) & \ge \inf_{\widehat{\theta}_n} \max_{v \in \pm 1} \mathbb{E}_{Q^n_{v\gamma}} \left[\mathsf{Risk}^\Theta_{Q_{v\gamma}}(\widehat{\theta}_n)\right] \\
& \stackrel{(i)}{\ge} \mathsf{sep}(\mathsf{Risk}_{Q_\gamma}, \mathsf{Risk}_{Q_{-\gamma}}, \Theta) \left(1 - \sqrt{\frac{n}{2} D_{\rm kl}\left(Q_\gamma \| Q_{-\gamma}\right)}\right) \\
& \stackrel{(ii)}{\ge} \gamma \, \mathsf{sep}(\mathsf{Risk}_{P_1}, \mathsf{Risk}_{P_{-1}}, \Theta) \left(1 - \sqrt{\frac{n\gamma}{2} D_{\rm kl}\left(P_1 \| P_{-1}\right)}\right) \\
& \stackrel{(iii)}{\ge} \frac{\gamma q_\ell^\star(t, y) |\ell'(t, y)| \delta}{2} \varepsilon \left(1 - \sqrt{n\gamma q_\ell^\star(t, y) \varepsilon^2 / 2}\right),
\end{aligned}$$

where inequality (i) is Le Cam's inequality (Lemma 8), inequality (ii) follows via Lemma 10, and Lemma 9 gives inequality (iii) whenever $\varepsilon \le \frac{3}{5}$. Choosing $\varepsilon^2 = \frac{1}{2n\gamma q_\ell^\star(t,y)}$ (where we use that n is large enough that $\varepsilon^2 \le \frac{1}{3}$) yields the lower bound

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \ge \frac{\sqrt{\gamma q_\ell^\star(t, y)}}{4\sqrt{n}} |\ell'(t, y)| \delta \tag{13}$$

valid for all $\delta \ge 0$ satisfying

$$\delta \le \frac{|\ell'(t, y)|}{\sup_{|\Delta| \le \delta} \ell''(t + \Delta, y)} \frac{\sqrt{q_\ell^\star(t, y)}}{\sqrt{2n\gamma}}.$$

This is circular, but we note that if we define

$$m_n(\delta) = m_n(\delta, t, y, \ell, \gamma) := \min\left\{\delta, \frac{|\ell'(t, y)|}{\sup_{|\Delta| \le \delta} \ell''(t + \Delta, y)} \frac{\sqrt{q_\ell^\star(t, y)}}{\sqrt{2n\gamma}}\right\}$$

then $m_n(\delta)$ satisfies $m_n(\delta) \le \frac{|\ell'(t,y)|}{\sup_{|\Delta| \le m_n(\delta)} \ell''(t+\Delta,y)} \frac{\sqrt{q_\ell^\star(t,y)}}{\sqrt{2n\gamma}}$, and substituting $m_n(\delta)$ for δ in the lower bound (13) gives

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \ge \frac{\sqrt{\gamma q_\ell^\star(t,y)}}{4\sqrt{n}} |\ell'(t,y)| \min\left\{\delta, \frac{|\ell'(t,y)|}{\sup_{|\Delta| \le \delta} \ell''(t+\Delta, y)} \frac{\sqrt{q_\ell^\star(t,y)}}{\sqrt{2n\gamma}}\right\},$$

valid for all $\delta \ge 0$ satisfying $2(t^2 + \delta^2) \le R^2 B^2$ as in Eq. (12).

A.2 Proof of Lemma 9

Recall throughout that $\varepsilon \in [0, 1]$. We provide the proof in two parts. In the first, we demonstrate the claimed risk separation by a Taylor approximation argument, and in the second, we provide the claimed bound on the KL divergence.

To show the risk separation, choose orthogonal vectors $v, w \in \mathbb{R}^d$ satisfying $\|v\|_2 = \|w\|_2 = R/\sqrt{2}$ and $\langle v, w \rangle = 0$, so that $\|v \pm w\|_2 = R$. For values $q \in [0, 1]$, $\alpha \in [-1, 1]$, and $y_0 \in \mathcal{Y}$ to be specified presently, we consider distributions on $\mathbb{R}^d \times \mathcal{Y}$ defined for $\sigma \in \{-1, 0, 1\}$ by

$$P_i : (X, Y) = \begin{cases} (\alpha v, y_0) & \text{with probability } 1 - q \\ (v + w, y) & \text{with probability } \frac{q}{2}(1 + \sigma\varepsilon) \\ (v - w, y) & \text{with probability } \frac{q}{2}(1 - \sigma\varepsilon). \end{cases} \tag{14}$$

In this case, the risk evidently satisfies

$$\mathsf{Risk}_{P_0}(\theta) = (1 - q)\ell(\alpha\theta^T v, y_0) + \frac{q}{2}\left[\ell(\theta^T(v + w), y) + \ell(\theta^T(v - w), y)\right].$$

We now construct its minimizer by judicious choice of q, where scaling by $\alpha \in [-1, 1]$ is sometimes necessary. Define $\theta_0 = \frac{2}{R^2} tv$, so that $\|\theta_0\|_2 = \sqrt{2}t/R \le B$, $\theta_0^T w = 0$ and $\theta_0^T v = t$, and

$$\nabla \mathsf{Risk}_{P_0}(\theta_0) = \alpha(1 - q)\ell'(\alpha t, y_0)v + q\ell'(t, y)v,$$

so that if

$$q = \frac{\alpha\ell'(\alpha t, y_0)}{\alpha\ell'(\alpha t, y_0) - \ell'(t, y)}$$

satisfies $q \in [0, 1]$, we have $\nabla \mathsf{Risk}_{P_0}(\theta_0) = 0$ and $\theta_0 \in \operatorname{argmin}_{\theta \in \Theta} \mathsf{Risk}_{P_0}(\theta)$. In particular, we may choose

$$q = q_\ell^\star(t, y) := \sup_{y_0 \in \mathcal{Y}, \alpha \in [-1,1]} \left\{\frac{\alpha\ell'(\alpha t, y_0)}{\alpha\ell'(\alpha t, y_0) - \ell'(t, y)} \text{ s.t. } \operatorname{sign}(\alpha\ell'(\alpha t, y_0)) \ne \operatorname{sign}(\ell'(t, y))\right\}.$$

We will perform a Taylor approximation of the risks Risk_{P_σ} for $\sigma \in \{\pm 1\}$ around θ_0 to show the desired separation bound. To that end, for $\delta \in \mathbb{R}$ define the shifted vector

$$\theta_\delta := \frac{2}{R^2}(tv + \delta w) = \theta_0 + \frac{2\delta}{R^2} w,$$

for which we have $\|\theta_\delta\|_2^2 = 2(t^2 + \delta^2)/R^2$ and $\theta_\delta^T(v \pm w) = t \pm \delta$. Using the risk expansion

$$\mathsf{Risk}_{P_\sigma}(\theta) = \mathsf{Risk}_{P_0}(\theta) + \frac{q\sigma\varepsilon}{2}\left[\ell(\theta^T(v + w), y) - \ell(\theta^T(v - w), y)\right] \tag{15}$$

and the Taylor approximation

$$\ell(t + \delta, y) = \ell(t, y) + \ell'(t, y)\delta + \frac{\delta^2}{2}\ell''(t + \Delta, y) \text{ for some } \Delta \in [0, \delta],$$

we obtain

$$\begin{aligned}\mathsf{Risk}_{P_\sigma}(\theta_\delta) &= (1-q)\ell(t,y_0) + q\ell(t,y) + \sigma\varepsilon q\ell'(t,y)\cdot\delta + \frac{\delta^2}{2}\mathrm{rem}(\delta) \\ &= \mathsf{Risk}_{P_0}(\theta_0) + \sigma\varepsilon q\ell'(t,y)\cdot\delta + \frac{\delta^2}{2}\mathrm{rem}(\delta),\end{aligned}$$

where the remainder term $|\mathrm{rem}(\delta)| \le \sup_{|\Delta|\le\delta} \ell''(t+\Delta, y)$. In particular, if $|\delta|$ is small enough that the conditions (12) hold, that is,

$$\sup_{|\Delta|\le|\delta|} |\delta|\ell''(t+\Delta,y) \le \varepsilon q|\ell'(t,y)| \quad \text{and} \quad 2(t^2+\delta^2) \le R^2B^2,$$

then setting $s = -\operatorname{sign}(\sigma\ell'(t,y))$ and letting $\delta \ge 0$ satisfy the conditions (12), we have $\theta_{s\delta} \in \Theta$ and

$$\inf_{\theta\in\Theta} \mathsf{Risk}_{P_\sigma}(\theta) \le \mathsf{Risk}_{P_\sigma}(\theta_{s\delta}) \le \mathsf{Risk}_{P_0}(\theta_0) - \frac{q\varepsilon}{2}|\ell'(t,y)|\delta.$$

Combining this inequality with the risk expansion (15), we see immediately that if $\theta \in \Theta$ satisfies $\ell(\theta^T(v+w),y) \ge \ell(\theta^T(v-w),y)$ then

$$\mathsf{Risk}_{P_1}(\theta) \ge \inf_{\theta\in\Theta} \mathsf{Risk}_{P_1}(\theta) + \frac{q\varepsilon}{2}|\ell'(t,y)|\delta,$$

and conversely $\ell(\theta^T(v-w),y) \le \ell(\theta^T(v-w),y)$ implies

$$\mathsf{Risk}_{P_{-1}}(\theta) \ge \inf_{\theta\in\Theta} \mathsf{Risk}_{P_1}(\theta) + \frac{q\varepsilon}{2}|\ell'(t,y)|\delta.$$

As θ_0 minimizes Risk_{P_0}, the expansion (15) implies that any θ minimizing $\mathsf{Risk}_{P_i}(\theta)$ over Θ necessarily satisfies $\sigma[\ell(\theta^T(v+w),y) - \ell(\theta^T(v-w),y)] < 0$, so we obtain the risk separation

$$\mathsf{sep}(\mathsf{Risk}^\Theta_{P_1}, \mathsf{Risk}^\Theta_{P_{-1}}, \Theta) \ge \frac{q\epsilon}{2}|\ell'(t,y)|\delta,$$

valid for any δ satisfying the constraints (12), which proves the claimed risk separation in the lemma.

To see the KL bound in Lemma 9, we note that for any pair of distributions of the form (14), we have

$$D_{\rm kl}\left(P_1 \| P_{-1}\right) = \frac{q(1+\varepsilon)}{2}\log\frac{1+\varepsilon}{1-\varepsilon} + \frac{q(1-\varepsilon)}{2}\log\frac{1-\varepsilon}{1+\varepsilon} = q\varepsilon\log\frac{1+\varepsilon}{1-\varepsilon} \overset{(\star)}{\le} q\varepsilon^2,$$

where inequality $(\star)$ is valid for $\varepsilon \le \frac{3}{5}$.

A.3 Proof of Lemma 10

Let P_0 have any distribution on $Y \mid X$ and $P_0(X = \mathbf{0}) = 1$, that is, the marginal over X is supported completely on $\mathbf{0}$. Then it is immediate that for $Q_{\pm\gamma} = (1-\gamma)P_0 + \gamma P_{\pm 1}$, we have

$$\mathsf{Risk}_{Q_{\pm\gamma}}(\theta) = (1-\gamma)\mathbb{E}_{P_0}[\ell(0,Y)] + \gamma\mathsf{Risk}_{P_{\pm1}}(\theta),$$

and therefore $\mathsf{Risk}^\Theta_{Q_{\pm\gamma}}(\theta) = \gamma\mathsf{Risk}^\Theta_{P_{\pm1}}(\theta)$. It is therefore immediate that $\mathsf{sep}(\mathsf{Risk}_{Q_\gamma}, \mathsf{Risk}_{Q_{-\gamma}}, \Theta) = \gamma\,\mathsf{sep}(\mathsf{Risk}_{P_1}, \mathsf{Risk}_{P_{-1}}, \Theta)$. For the gap on the KL divergence, we use joint convexity to obtain

$$D_{\rm kl}\left(Q_\gamma \| Q_{-\gamma}\right) = D_{\rm kl}\left((1-\gamma)P_0 + \gamma P_1 \| (1-\gamma)P_0 + \gamma P_{-1}\right) \le (1-\gamma)\underbrace{D_{\rm kl}\left(P_0 \| P_0\right)}_{=0} + \gamma D_{\rm kl}\left(P_1 \| P_{-1}\right).$$

A.4 Proof of Proposition 2

By assumption, there exist t, y, y_0 satisfying $\ell'(t,y)\ell'(t,y_0) < 0$, and $\ell''(t,y) = 0$. Then it is evidently the case that $q_\ell^\star \equiv q_\ell^\star(t,y) > 0$, so that we obtain the lower bound

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c \sup_{0 \leq \delta \leq RB/2} |\ell'(t,y)| \min \left\{ \delta \sqrt{\gamma q_\ell^\star}, \frac{|\ell'(t,y)|}{\sup_{|\Delta| \leq \delta} \ell''(t+\Delta, y)} \frac{q_\ell^\star}{\sqrt{2n}} \right\}.$$

Now, we recall that ℓ is $\mathcal{C}^3$ near t and by assumption $\ell''(t,y) = 0$, for all suitably small δ we obtain $|\ell''(t+\Delta, y)| \geq |\ell'''(t,y)||\Delta|/2$, and so in particular for all small δ,

$$\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c|\ell'(t,y)| \min \left\{ \delta \sqrt{\gamma q_\ell^\star}, \frac{2|\ell'(t,y)|}{|\ell'''(t,y)|\delta} \frac{q_\ell^\star}{\sqrt{2n}} \right\}.$$

Set $\delta^2 = \frac{1}{\sqrt{n}}$ to obtain that for some problem-dependent constant $c_{\rm prob}$, we have $\mathsf{Lin}(\ell, \mathcal{Y}, R, B, n, \gamma) \geq c_{\rm prob} \frac{1}{n^{1/4}}$. Substitute this lower bound in Theorem 1.

B Technical appendices

B.1 Proofs of mixability in Table 1

We assume that $\mathcal{Y}$ is discrete and of size k (it is not difficult to obtain a result when $\mathcal{Y} = \mathbb{N}$), so that we may identify distributions on $\mathcal{Y}$ with vectors $p \in \Delta_k := \{v \in R_+^k \mid \mathbb{1}^T v = 1\}$, the probability simplex in $\mathbb{R}^k$. Consider any $\mathcal{C}^2$ function $h : \Delta \to \mathbb{R}$, noting that

$$\begin{aligned} \nabla \exp(-\eta h(p)) &= -\eta \exp(-\eta h(p)) \nabla h(p), \\ \nabla^2 \exp(-\eta h(p)) &= \eta \exp(-\eta h(p)) \left[\eta \nabla h(p) \nabla h(p)^T - \nabla^2 h(p) \right]. \end{aligned}$$

We consider each of the columns of the table in turn. Thus to demonstrate exp-concavity it is sufficient that $\nabla^2 h(p) \succeq \eta \nabla h(p) \nabla h(p)^T$ for all $p \in \Delta_k$.

1. For $L_{\log}$, we take $h(p) = -\log p$, for which it is immediate that $\eta = 1$ suffices as $\exp(-h(p)) = p$.
2. For $L_{\rm sq}$, we have $h(p) = \frac{1}{2}(p-1)^2$, $h'(p) = (p-1)$, and $h''(p) = 1$, so $\eta = 1$ suffices.
3. For $L_{\rm hel}$, we have $h(p) = (\sqrt{p} - 1)^2 = p - 2\sqrt{p} + 1$, $h'(p) = 1 - \frac{1}{\sqrt{p}}$, and $h''(p) = \frac{1}{2p^{3/2}}$. Thus, we seek η such that
$$\frac{1}{2p^{3/2}} \geq \eta(1 - 1/\sqrt{p})^2 \quad \text{or} \quad \frac{1}{2} \geq \eta(p^{3/2} - 2p + \sqrt{p})$$
for all $p \in [0,1]$. Letting $\beta = \sqrt{p}$ and solving for the stationary points of $\beta^3 - 2\beta + \beta$ at $\sqrt{p} = \beta = 1/3$ and $\beta = 1$, we see it is sufficient that $1 \geq 2\eta(1/27 - 2/9 + 1/3) = \frac{8}{27}\eta$, or $\eta \leq \frac{27}{8}$.
4. For $L_{\rm quad}$, we have $h(p) = \frac{1}{2}\|p - e_y\|_2^2$, so it suffices that $I - \eta(p - e_y)(p - e_y)^T \succeq 0$, or $\eta \leq \frac{1}{2}$.

B.2 Proof of Theorem 5

Recall Definition A.1 of the separation between two functions. We first recall the essentially standard reduction of estimation to testing, which proceeds as follows. Let $\mathcal{V}$ be a finite set indexing a collection $\{P_v\}_{v \in \mathcal{V}}$ of distributions on $\mathcal{X} \times \mathcal{Y}$ and a collection of functions $\{f_v\}$. Consider the following process: draw $V \in \mathcal{V}$ uniformly at random, and conditional on $V = v$, observe $(X_i, Y_i) \stackrel{\rm iid}{\sim} P_v$ for $i = 1, 2, \ldots, n$. Then we have the following lemma, which reduces optimization of f_v to testing the index V (see, e.g. [12, Sec. 5] or [41, Ch. 15]).

Lemma 11. *Let $f_v^\star = \inf_{\theta\in\Theta} f_v(\theta)$ for $v \in \mathcal{V}$. Then*

$$\inf_{\widehat{\theta}_n} \max_{v\in\mathcal{V}} \mathbb{E}_{P_v^n}\left[f_v(\widehat{\theta}_n(X_1^n, Y_1^n)) - f_v^\star\right] \geq \min_{v\neq w\in\mathcal{V}} \mathsf{sep}(f_v, f_w, \Theta) \cdot \inf_{\widehat{\Psi}_n} \mathbb{P}(\widehat{\Psi}_n(X_1^n, Y_1^n) \neq V),$$

where the infima are over procedures $\widehat{\theta}_n : \mathcal{X}^n \times \mathcal{Y}^n \to \Theta$ and all measurable functions $\widehat{\Psi}_n$, respectively.

We thus lower bound the probability of error in testing, $\widehat{\Psi} \neq V$, for which we use Fano's inequality [9]:

Lemma 12 (Fano's Inequality)**.** *Let $I(V; X_1^n, Y_1^n)$ be the (Shannon) mutual information between V and (X_1^n, Y_1^n), where $(X_i, Y_i) \stackrel{\text{iid}}{\sim} P_v$ conditional on $V = v$ and V is uniform on $\mathcal{V}$. Then for any $\widehat{\Psi}$,*

$$\mathbb{P}(\widehat{\Psi}(X_1^n, Y_1^n) \neq V) \geq 1 - \frac{I(V; X_1^n, Y_1^n) + \log 2}{\log|\mathcal{V}|}.$$

Now, we define the collection of problems we consider and their induced risks. Let X be uniform on $\{\pm 1\}^d$, and let

$$p_\theta(y \mid x) = \exp(y\theta^T x - A(\theta^T x))$$

be the density of P_θ with respect to the base measure ν. For a value $\delta \geq 0$ to be chosen, let P_v be the joint distribution on (X, Y) with $\theta = \delta v$. We first demonstrate that these induce a separation in the expected log loss of a predictive distribution $p(\cdot \mid x)$, where for such a p we define the risk

$$\mathsf{Risk}_{\delta v}(p) := \mathbb{E}_{P_v}\left[L_{\log}(p(\cdot \mid X), Y)\right] = \mathbb{E}_{P_v}[-\log p(Y \mid X)],$$

where we note that $p_{\delta v}$ minimizes $\mathsf{Risk}_{\delta v}$ as it is well-specified. The key to applying Lemmas 11 and 12 are the following two technical results, which respectively lower bound the separation and upper bound the KL-divergence between distributions. We defer proofs to Sections B.2.1 and B.2.2.

Lemma 13. *Let $\mathcal{P}$ be the collection of all conditional probability distributions on $Y \mid X$. There exists a constant $C(A)$ depending only on the log partition function $A(\cdot)$ such that for all $\delta \geq 0$ and $u, v \in \mathbb{R}^d$,*

$$\mathsf{sep}(\mathsf{Risk}_{\delta v}, \mathsf{Risk}_{\delta w}, \mathcal{P}) \geq \frac{1}{16} A''(0)\delta^2 \left\|v - w\right\|_2^2 - C(A)\delta^3 d^{3/2} \max\{\left\|v\right\|_2, \left\|w\right\|_2\}^3.$$

Lemma 14. *For $v \in \mathbb{R}^d$, let $P_{\delta v}$ denote the joint distribution over $X \sim \mathsf{Uni}(\{-1, 1\}^d)$ and $Y \mid X = x$ having exponential family density $p_{\delta v}(y \mid x) = \exp(y\theta^T x - A(\theta^T x))$. There exists a constant $C(A)$ depending only on the log partition function $A(\cdot)$ such that for all $\delta \in [0, 1]$ and u, v satisfying $\left\|u\right\|_2 \leq 1$, $\left\|v\right\|_2 \leq 1$,*

$$D_{\rm kl}\left(P_{\delta v} \| P_{\delta w}\right) \leq \frac{\delta^2}{2} A''(0) \left\|v - w\right\|_2^2 + C(A)\delta^3 \left\|v - w\right\|_2^3.$$

With these two lemmas, the result is relatively straightforward. We consider two cases: that $d \geq 8$ and (for completeness) that $d \leq 8$, which we defer temporarily. Let $d \geq 8$. By a standard volume argument [41, Ch. 15], there exists a packing set $\mathcal{V} \subset \{v \in \mathbb{R}^d \mid \left\|v\right\|_2 = 1\}$ of the ℓ_2 sphere satisfying $|\mathcal{V}| \geq \exp(d/4)$ and $\left\|v - w\right\|_2 \geq \frac{1}{2}$ for each $v \neq w \in \mathcal{V}$. Let V be uniform on $\mathcal{V}$ as in our construction above. Then naive bounds on the mutual information $I(V; X_1^n, Y_1^n)$ yield that

$$I(V; X_1^n, Y_1^n) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,w\in\mathcal{V}} D_{\rm kl}\left(P_v^n \| P_w^n\right) \stackrel{(\star)}{\leq} n \cdot \frac{1}{|\mathcal{V}|^2} \sum_{v,w\in\mathcal{V}} \delta^2 A''(0) \left\|v - w\right\|_2^2 \leq 4n\delta^2 A''(0),$$

where inequality $(\star)$ holds for any sufficiently small $\delta \geq 0$ by Lemma 14. Applying Lemmas 11 and 13 by noting that $\left\|v - w\right\|_2 \geq \frac{1}{2}$, there exists a numerical constant $c > 0$ such that for small enough $\delta \geq 0$,

$$\begin{aligned}\mathfrak{M}_n(\Theta, \mathcal{P}, 0) &\geq cA''(0)\delta^2 \inf_{\widehat{\Psi}_n} \mathbb{P}(\widehat{\Psi}_n(X_1^n, Y_1^n) \neq V) \\ &\geq cA''(0)\delta^2 \left(1 - \frac{I(V; X_1^n, Y_1^n) + \log 2}{\log|\mathcal{V}|}\right),\end{aligned}$$

where the second inequality is Fano's inequality (Lemma 12). Applying the preceding bound on the mutual information and that $\log|\mathcal{V}| \ge d/4$ then implies

$$\mathfrak{M}_n(\Theta, \mathcal{P}, 0) \ge cA''(0)\delta^2 \left(1 - \frac{16n\delta^2 A''(0) + 4\log 2}{d}\right).$$

Choosing $\delta^2 = \frac{d}{32A''(0)n}$ then gives the theorem in the case that $d \ge 8$.

For the final case that $d \le 8$, we apply Le Cam's method as in our proof of Theorem 1. We assume that $d = 1$, as increasing the dimension simply increases the risk bound, and let $X \sim \mathsf{Uni}(\{-1, 1\})$. Recalling Lemma 8, we apply Lemmas 13 and 14 to obtain

$$\mathfrak{M}_n(\Theta, \mathcal{P}, 0) \ge cA''(0)\delta^2 \left(1 - \sqrt{Cn\delta^2 A''(0)}\right),$$

where $0 < c$ and $C < \infty$ are numerical constants. Setting $\delta^2 = \frac{1}{4CnA''(0)}$ then yields the result.

B.2.1 Proof of Lemma 13

We define the excess risk functional

$$f_{\delta v}(p) := \mathsf{Risk}_{\delta v}(p) - \inf_p \mathsf{Risk}_{\delta v}(p) = \mathbb{E}_{P_v}\left[\log \frac{p_{\delta v}(Y \mid X)}{p(Y \mid X)}\right] = \mathbb{E}\left[D_{\rm kl}\left(p_{\delta v}(\cdot \mid X) \| p(\cdot \mid X)\right)\right],$$

where we have used that the exponential family model $p_{\delta v}$ minimizes $\mathsf{Risk}_{\delta v}$, and we note that

$$\mathsf{sep}(f_{\delta v}, f_{\delta w}, \mathcal{P}) \ge \frac{1}{2} \inf_{p \in \mathcal{P}} \{f_{\delta v}(p) + f_{\delta w}(p)\}$$

(this inequality is valid for any functions and set $\mathcal{P}$). Thus

$$2\,\mathsf{sep}(\mathsf{Risk}_{\delta v}, \mathsf{Risk}_{\delta w}, \mathcal{P}) = 2\,\mathsf{sep}(f_{\delta v}, f_{\delta w}, \mathcal{P}) \ge \mathbb{E}\left[\inf_p \left\{D_{\rm kl}\left(p_{\delta v}(\cdot \mid X) \| p\right) + D_{\rm kl}\left(p_{\delta w}(\cdot \mid X) \| p\right)\right\}\right].$$

Now we use that for any three distributions P_0, P_1, Q, if $\overline{P} = \frac{1}{2}(P_0 + P_1)$ then

$$D_{\rm kl}\left(P_0 \| Q\right) + D_{\rm kl}\left(P_1 \| Q\right) = D_{\rm kl}\left(P_0 \| \overline{P}\right) + D_{\rm kl}\left(P_1 \| \overline{P}\right) + 2D_{\rm kl}\left(\overline{P} \| Q\right) \ge D_{\rm kl}\left(P_0 \| \overline{P}\right) + D_{\rm kl}\left(P_1 \| \overline{P}\right),$$

and substituting this into the preceding lower bound on the separation gives

$$\begin{aligned} 2\,\mathsf{sep}(\mathsf{Risk}_{\delta v}, \mathsf{Risk}_{\delta w}, \mathcal{P}) \ge{}& \mathbb{E}\left[D_{\rm kl}\left(p_{\delta v}(\cdot \mid X) \| (1/2)(p_{\delta v}(\cdot \mid X) + p_{\delta w}(\cdot \mid X))\right)\right] \\ &+ \mathbb{E}\left[D_{\rm kl}\left(p_{\delta w}(\cdot \mid X) \| (1/2)(p_{\delta v}(\cdot \mid X) + p_{\delta w}(\cdot \mid X))\right)\right], \end{aligned} \tag{16}$$

where the outer expectation is over $X \sim \mathsf{Uni}(\{-1, 1\}^d)$.

We now provide an asymptotic lower bound on the KL divergences, focusing on a single term given $X = x$ in the lower bound (16). By a Taylor expansion,

$$\log(1 + e^t) = \log 2 + \frac{t}{2} + \frac{t^2}{8} \pm O(1)t^3,$$

where $O(1)$ denotes a universal numerical constant and the expansion is valid for all $t \in \mathbb{R}$ because $t \mapsto \log(1 + e^t)$ is 1-Lipschitz. Using the shorthand $t = \delta v^T x$ and $u = \delta w^T x$ and $p_t(y) = p_{\delta v}(\cdot \mid x)$ and similarly for p_u, we have

$$\begin{aligned} &D_{\rm kl}\left(p_t \| (1/2)(p_t + p_u)\right) = \int p_t(y) \log \frac{2}{1 + p_u(y)/p_t(y)} d\nu \\ &= \int p_t(y) \left[\log 2 - \log\left(1 + e^{y(u-t) - (A(u) - A(t))}\right)\right] d\nu \\ &= \int p_t(y) \left[\frac{y(t-u) + A(u) - A(t)}{2} - \frac{\left(y(t-u) + A(u) - A(t)\right)^2}{8} \pm O(1)\left(y(t-u) + A(u) - A(t)\right)^3\right] d\nu. \end{aligned}$$

By standard properties of exponential families, if $\mathbb{E}_t$ denotes expectation under p_t, we have $A'(t) = \mathbb{E}_t[Y]$, and A is $\mathcal{C}^\infty$ near 0, so that $A(u) - A(t) = A'(t)(u-t) + \frac{1}{2}(u-t)^2 A''(\widetilde{u})$ for some $\widetilde{u} \in [u,t]$. We may thus write

$$\begin{aligned}
&D_{\rm kl}\left(p_t \| (1/2)(p_t + p_u)\right) \\
&= \int p_t(y) \Bigg[\frac{(y - A'(t))(t-u)}{2} + \frac{(u-t)^2 A''(\widetilde{u})}{4} - \frac{\left((y-A'(t))(t-u) + (u-t)^2 A''(\widetilde{u})/2\right)^2}{8} \\
&\qquad \pm O(1)\left[|y - A'(t)|^3 |t-u|^3 + (t-u)^6 A''(\widetilde{u})^3\right] \Bigg] d\nu \\
&= \frac{1}{4} A''(\widetilde{u})(u-t)^2 - \frac{1}{8} A''(t)(u-t)^2 - \frac{1}{32} A''(\widetilde{u})^2 (u-t)^4 \pm O(1) \mathbb{E}_t[|Y - \mathbb{E}_t[Y]|^3] |t-u|^3.
\end{aligned}$$

As $A(\cdot)$ exists in a neighborhood of 0, the moment generating functions of p_t, p_u exist, this expansion is uniform in u, t near 0, and so we obtain

$$D_{\rm kl}\left(p_t \| (1/2)(p_t + p_u)\right) = \frac{1}{8}(u-t)^2 A''(0) \pm C(A)|u-t|^3, \tag{17}$$

where $C(A)$ is a constant depending on the log partition function $A(\cdot)$, and the expansion is uniform for u, t in a neighborhood of 0.

Finally, we recall that $t = \delta v^T x$ and $u = \delta w^T x$, and as $|v^T x| \le \|v\|_2 \|x\|_2$, we have the lower bound

$$\inf_p \left\{ D_{\rm kl}\left(p_{\delta v}(\cdot \mid x) \| p\right) + D_{\rm kl}\left(p_{\delta w}(\cdot \mid x) \| p\right) \right\} \ge \frac{1}{8} A''(0) \delta^2 (x^T(w-v))^2 - C(A)\delta^3 d^{3/2} \max\{\|w\|_2, \|v\|_2\}^3.$$

Substituting this into our lower bound (16) and using that $\mathbb{E}[XX^T] = I_d$ by construction then gives the lemma.

B.2.2 Proof of Lemma 14

Without loss of generality, assume that $\|v\|_2 \ge \|w\|_2$. We have

$$D_{\rm kl}\left(P_{\delta v} \| P_{\delta w}\right) = \mathbb{E}\left[D_{\rm kl}\left(p_{\delta v}(\cdot \mid X) \| p_{\delta w}(\cdot \mid X)\right)\right].$$

Fix x temporarily, and consider the inner KL-divergence term. As in the proof of Lemma 13, we use the shorthands $t = \delta v^T x$, $u = \delta w^T x$, $p_t = p_{\delta v}(\cdot \mid x)$ and $p_u = p_{\delta w}(\cdot \mid x)$, noting that $|t| \le \delta\sqrt{d}\|v\|_2$ and similarly for u. Then writing $\mathbb{E}_t$ for expectation under p_t, we have

$$D_{\rm kl}\left(p_t \| p_u\right) = \mathbb{E}_t\left[Y(t-u)\right] + A(u) - A(t) = A(u) - A(t) - A'(t)(u-t) = \frac{1}{2} A''(\widetilde{u})(u-t)^2,$$

where $\widetilde{u} \in [u,t]$. As A is $\mathcal{C}^\infty$ near 0, we obtain that for a constant $C(A)$ depending only on A that

$$D_{\rm kl}\left(p_t \| p_u\right) \le \frac{1}{2} A''(0)(u-t)^2 + C(A)|u-t|^3,$$

valid for all $u, t \in [-\delta\sqrt{d}\|v\|_2, \delta\sqrt{d}\|v\|_2]$. We we obtain

$$D_{\rm kl}\left(P_{\delta v} \| P_{\delta w}\right) \le \frac{\delta^2}{2} A''(0) \mathbb{E}[(X^T(v-w))^2] + C(A)\delta^3 \mathbb{E}[|X^T(v-w)|^3],$$

and using $\mathbb{E}[(X^T(v-w))^2] = \|v-w\|_2^2$ and $\mathbb{E}[|X^T v|^3] \lesssim \|v\|_2^3$ for $X \sim \mathsf{Uni}(\{-1,1\}^d)$ gives the lemma.

B.3 Proof of Theorem 6

Recall $\widehat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} \mathsf{Risk}_n(\theta)$ and Definition 3.1. For shorthand, we use the standard empirical process notation that $Pf = \mathbb{E}_P[f]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$. Let $\delta_n > 0$ be any sequence satisfying

$$\frac{\log\log n}{n} \ll \delta_n^2 \ll \frac{1}{\sqrt{n}}.$$

We will define a "good event," which is roughly that the empirical risk Risk_n approximates the true risk Risk_P well and a local quadratic approximation to both is accurate, and perform our analysis (essentially) conditional on this good event. To that end, let $\lambda_{\min} = \lambda_{\min}(\nabla^2\mathsf{Risk}_P(\theta^\star)$ and $\lambda_{\max} = \lambda_{\max}(\nabla^2\mathsf{Risk}_P(\theta^\star))$ be the minimum and maximum eigenvalues of $\nabla^2\mathsf{Risk}_P(\theta^\star)$. Recall that $\epsilon_1 > 0$ is the radius of the ball on which $\nabla^2 L(p_\theta(\cdot \mid x), y)$ is $M_{\text{Lip},2}(x,y)$ Lipschitz (Def. 3.1), and for an $\epsilon > 0$ to be determined

$$\mathcal{E}_n := \left\{P_n M_{\text{Lip},2} \le 2PM_{\text{Lip},2},\ \frac{\lambda_{\min}}{2} I \preceq \nabla^2\mathsf{Risk}_n(\theta) \preceq 2\lambda_{\max} I \text{ for } \|\theta - \theta^\star\|_2 \le \epsilon,\ \|\widehat{\theta}_n - \theta^\star\|_2 \le \delta_n\right\}. \tag{18}$$

We prove the theorem in a series of lemmas. The first shows that $\mathcal{E}_n$ occurs eventually, and the remainder we will demonstrate hold on the event.

Lemma 15. *For all sufficiently small $\epsilon > 0$, $\mathcal{E}_n$ happens eventually. That is, there is a (random) N, finite with probability 1, such that $\mathcal{E}_n$ occurs for all $n \ge N$.*

Proof. By the strong law of large numbers, we have $P_n M_{\text{Lip},2} \stackrel{a.s.}{\to} PM_{\text{Lip},2}$, so that $P_n M_{\text{Lip},2} \le 2PM_{\text{Lip},2}$ eventually, while Definition 3.1 implies that $\left\|\nabla^2\mathsf{Risk}_n(\theta) - \nabla^2\mathsf{Risk}_n(\theta^\star)\right\|_{\text{op}} \le 2PM_{\text{Lip},2}\epsilon$ for all $\|\theta - \theta^\star\|_2 \le \epsilon_1$ on the same event. Whenever ϵ is small enough that $PM_{\text{Lip},2}\epsilon \le \frac{\lambda_{\max}}{2}$ and $PM_{\text{Lip},2}\epsilon \le \frac{\lambda_{\min}}{4}$, we then obtain that $\frac{\lambda_{\min}}{2} I \preceq \nabla^2\mathsf{Risk}_n(\theta) \preceq 2\lambda_{\max} I$ by choosing

$$\epsilon \le \min\left\{\epsilon_1, \frac{\lambda_{\min}}{4PM_{\text{Lip},2}}\right\}.$$

Finally, we argue that $\|\widehat{\theta}_n - \theta^\star\|_2 \le \delta_n$ eventually. A standard argument [36, Thm. 5.7] and the Glivenko Cantelli theorem, which implies $\sup_{\theta\in\Theta} |\mathsf{Risk}_n(\theta) - \mathsf{Risk}_P(\theta)| \stackrel{a.s.}{\to} 0$ by the compactness of Θ, gives the consistency $\widehat{\theta}_n \stackrel{a.s.}{\to} \theta^\star$. As $\theta^\star \in \operatorname{int}\Theta$, Taylor's theorem implies that

$$0 = \nabla\mathsf{Risk}_n(\widehat{\theta}_n) = \nabla\mathsf{Risk}_n(\theta^\star) + (\nabla^2\mathsf{Risk}_n(\theta^\star) + E_n(\widehat{\theta}_n, \theta^\star))(\widehat{\theta}_n - \theta^\star),$$

where E_n is an error matrix that Definition 3.1 implies satisfies

$$\left\|E_n\right\|_{\text{op}} \le \frac{1}{n}\sum_{i=1}^n M_{\text{Lip},2}(X_i, Y_i)\|\widehat{\theta}_n - \theta^\star\|_2.$$

Thus $\left\|E_n\right\|_{\text{op}} \stackrel{a.s.}{\to} 0$, and as $\nabla^2\mathsf{Risk}_n(\theta^\star) \stackrel{a.s.}{\to} \nabla^2\mathsf{Risk}(\theta^\star)$, we have $\widehat{\theta}_n - \theta^\star = -(\nabla^2\mathsf{Risk}(\theta^\star) + E_n')^{-1}\nabla\mathsf{Risk}_n(\theta^\star)$, where $E_n' \stackrel{a.s.}{\to} 0$ is an error matrix. By the a.s. convergence $E_n' \to 0$ and law of the iterated logarithm,

$$\begin{aligned}\limsup_n \sqrt{\frac{n}{\log\log n}} &\left\|(\nabla^2\mathsf{Risk}(\theta^\star) + E_n')^{-1}\nabla\mathsf{Risk}_n(\theta^\star)\right\|_2 \\ &\le \left\|\nabla^2\mathsf{Risk}(\theta^\star)^{-1}\right\|_{\text{op}} \limsup_n \sqrt{\frac{n}{\log\log n}} \left\|\nabla\mathsf{Risk}_n(\theta^\star)\right\|_2 < \infty\end{aligned}$$

with probability 1. In particular, whenever $\delta_n^2 \gg \frac{\log\log n}{n}$, we have $\|\widehat{\theta}_n - \theta^\star\|_2 \le \delta_n$ eventually. □

An immediate consequence of the identifiability condition (iii) in Definition 3.1 and Taylor's theorem is the following lemma.

Lemma 16. *For all large enough n, on event $\mathcal{E}_n$ we have*

$$\mathsf{Risk}_n(\theta) \le \mathsf{Risk}_n(\widehat{\theta}_n) + 2\lambda_{\max}\|\theta - \widehat{\theta}_n\|_2^2 \;\; \textit{for all } \|\theta - \widehat{\theta}_n\|_2 \le \delta_n$$

and

$$\mathsf{Risk}_n(\theta) \ge \mathsf{Risk}_n(\widehat{\theta}_n) + \frac{1}{4}\lambda_{\min}\delta_n^2 \;\; \textit{for all } \theta \in \Theta \textit{ s.t. } \|\theta - \widehat{\theta}_n\|_2 \ge \delta_n.$$

Finally, we show that on $\mathcal{E}_n$ we have

$$\left\|\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - \mathsf{N}\Big(\widehat{\theta}_n, \frac{1}{n}\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)^{-1}\Big)\right\|_{\mathrm{TV}} \to 0.$$

For shorthand, let π_n be the probability distribution $\mathsf{N}(\widehat{\theta}_n, \frac{1}{n}\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)^{-1})$. We split the variation distance into two terms. Let $B_n = \delta_n\mathbb{B}_2^d$ be an ℓ_2 ball of radius δ_n. Then

$$2\left\|\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - \pi_n\right\|_{\mathrm{TV}} = \underbrace{\int_{\widehat{\theta}_n + B_n} |d\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - d\pi_n|}_{=:T_1} + \underbrace{\int_{\Theta\setminus\{\widehat{\theta}_n + B_n\}} |d\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - d\pi_n|}_{=:T_2} + \underbrace{\pi_n(\Theta^c)}_{=:T_3}. \tag{19}$$

We bound each of the terms T_i in turn. For the second term, we compute bounds on the densities themselves. Let $\theta \in \Theta \setminus \{\widehat{\theta}_n + B_n\}$. Then for any $c > 0$ small enough that $\frac{\lambda_{\min}}{4} - 2c^2\lambda_{\max} =: K > 0$,

$$\begin{aligned}\frac{d}{d\theta}\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}}(\theta) = \frac{\exp(-n\mathsf{Risk}_n(\theta))}{\int_\Theta \exp(-n\mathsf{Risk}_n(\theta'))d\theta'} &\le \frac{\exp(-n\mathsf{Risk}_n(\theta))}{\int_{\widehat{\theta}_n + cB_n}\exp(-n\mathsf{Risk}_n(\theta'))d\theta'} \\ &\overset{(i)}{\le} \frac{\exp(-\frac{\lambda_{\min}}{4}n\delta_n^2)}{\exp(-2\lambda_{\max}c^2\delta_n^2)\,\mathrm{Vol}(cB_n)} = \exp\left(-nK\delta_n^2 + d\log\frac{1}{c\delta_n} - c_d\right),\end{aligned}$$

where inequality (i) follows from Lemma 16 and $c_d = \log\mathrm{Vol}(\mathbb{B}_2^d)$ is the log volume of the ℓ_2-ball. A completely analogous calculation gives

$$\begin{aligned}\frac{d}{d\theta}\pi_n(\theta) &= \frac{\exp(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n))}{\int\exp(-\frac{n}{2}(\theta' - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta' - \widehat{\theta}_n))d\theta'} \\ &\le \frac{\exp(-\frac{\lambda_{\min}}{4}n\delta_n^2)}{\exp(-2\lambda_{\max}c^2\delta_n^2)\,\mathrm{Vol}(cB_n)} = \exp\left(-nK\delta_n^2 + d\log\frac{1}{c\delta_n} - c_d\right),\end{aligned}$$

where the inequality uses the definition (18) of $\mathcal{E}_n$. In particular, setting the constant $c = \frac{1}{4}\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}$, the term $K = \frac{1}{8}\lambda_{\min}$ and we may bound term T_2 in expression (19) by

$$T_2 \le 2\,\mathrm{Vol}(\Theta)\exp\left(-nK\delta_n^2 + \frac{d}{2}\log\frac{16\lambda_{\max}}{\lambda_{\min}\delta_n^2} - c_d\right) \to 0,$$

as $\delta_n \gg \frac{1}{\sqrt{n}}$ and $\delta_n \to 0$.

Let us turn to term T_1 in expression (19). For sets $A \subset \mathbb{R}^d$ we define the normalizing constants

$$Z_{A,n}^{\mathsf{N}} := \int_A \exp\Big(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)\Big)\,d\theta \;\text{ and }\; Z_{A,n}^{\mathsf{Vovk}} := \int_A \exp\Big(-n(\mathsf{Risk}_n(\theta) - \mathsf{Risk}_n(\widehat{\theta}_n))\Big)\,d\theta.$$

Changing notation slightly to let $B_n = \widehat{\theta}_n + \delta_n \mathbb{B}_2^d$, Lemma 16 implies the inequalies

$$\max\left\{Z^{\mathsf{Vovk}}_{\Theta\setminus B_n,n}, Z^{\mathsf{N}}_{\Theta\setminus B_n,n}\right\} \le \mathrm{Vol}(\Theta)\exp\left(-\frac{\lambda_{\min}}{4}n\delta_n^2\right)$$

and

$$\min\left\{Z^{\mathsf{N}}_{B_n,n}, Z^{\mathsf{Vovk}}_{B_n,n}\right\} \ge \exp(-2n\lambda_{\max}c^2\delta_n^2)\,\mathrm{Vol}(c\delta_n\mathbb{B}_2^d),$$

valid for any $c \le 1$. Thus, the ratio

$$\begin{aligned}\rho_n := \max\left\{\frac{Z^{\mathsf{Vovk}}_{\Theta\setminus B_n,n}}{Z^{\mathsf{N}}_{B_n,n}}, \frac{Z^{\mathsf{N}}_{\Theta\setminus B_n,n}}{Z^{\mathsf{Vovk}}_{B_n,n}}\right\} &\le \frac{\mathrm{Vol}(\Theta)\exp(-\frac{\lambda_{\min}}{4}n\delta_n^2)}{\exp(-2n\lambda_{\max}c^2\delta_n^2)\,\mathrm{Vol}(c\delta_n\mathbb{B}_2^d)} \\ &\le \mathrm{Vol}(\Theta)\exp\left(-n\delta_n^2\left(\frac{\lambda_{\min}}{4} - 2c^2\lambda_{\max}\right) + d\log\frac{1}{c\delta_n} - c_d\right),\end{aligned}$$

where as before $c_d = \log\mathrm{Vol}(\mathbb{B}_2^d)$, so that for all small $c > 0$ we have $\rho_n \to 0$ as $n \to \infty$ on the event $\mathcal{E}_n$. We may then bound the normalizing constant ratio by

$$\frac{Z^{\mathsf{Vovk}}_{B_n,n}}{Z^{\mathsf{N}}_{B_n,n}} + \rho_n \ge \frac{Z^{\mathsf{Vovk}}_{B_n,n} + Z^{\mathsf{Vovk}}_{\Theta\setminus B_n,n}}{Z^{\mathsf{N}}_{B_n,n}} \ge \frac{Z^{\mathsf{Vovk}}_{\Theta,n}}{Z^{\mathsf{N}}_{\Theta,n}} \ge \frac{Z^{\mathsf{Vovk}}_{B_n,n}}{Z^{\mathsf{N}}_{B_n,n} + Z^{\mathsf{N}}_{\Theta\setminus B_n,n}} \ge \left(\frac{Z^{\mathsf{N}}_{B_n,n}}{Z^{\mathsf{Vovk}}_{B_n,n}} + \rho_n\right)^{-1}. \tag{20}$$

Performing a Taylor expansion, on $\mathcal{E}_n$, for any $\theta \in B_n$ the Lipschitz continuity of $\nabla^2\mathsf{Risk}_n(\theta)$ implies

$$\mathsf{Risk}_n(\theta) = \mathsf{Risk}_n(\widehat{\theta}_n) + \frac{1}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n) \pm PM_{\mathrm{Lip},2}\cdot\delta_n^3.$$

Using this $O(\delta_n^3)$ remainder term, we then immediately obtain the ratio bounds

$$\frac{Z^{\mathsf{Vovk}}_{B_n,n}}{Z^{\mathsf{N}}_{B_n,n}} = \frac{\int_{B_n}\exp\left(-n(\mathsf{Risk}_n(\theta) - \mathsf{Risk}_n(\widehat{\theta}_n))\right)d\theta}{\int_{B_n}\exp(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n))d\theta} \in \exp\left(\pm PM_{\mathrm{Lip},2}\cdot\delta_n^3\right).$$

Substituting this containment in the inequalities (20), we find that for all large enough n, on the event $\mathcal{E}_n$ in Eq. (18), we have the bounds

$$\exp(-PM_{\mathrm{Lip},2}\delta_n^3) - O(1)\rho_n \le \frac{Z^{\mathsf{Vovk}}_{\Theta,n}}{Z^{\mathsf{N}}_{\Theta,n}} \le \exp(PM_{\mathrm{Lip},2}\cdot\delta_n^3) + O(1)\rho_n. \tag{21}$$

Finally, we return to computing the densities in the term T_1 in Eq. (19). Let $Z_n^{\mathsf{N}} = Z^{\mathsf{N}}_{\mathbb{R}^d,n}$, where an argument similar to those above shows that $Z_n^{\mathsf{N}}/Z^{\mathsf{N}}_{\Theta,n} \to 1$ as $n \to \infty$. Defining the remainder $\mathrm{rem}_n(\theta) = \mathsf{Risk}_n(\theta) - \mathsf{Risk}_n(\widehat{\theta}_n) - \frac{1}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)$ and using that $\|\mathrm{rem}_n(\theta)\|_2 \le PM_{\mathrm{Lip},2}\cdot\delta_n^3$ for any $\theta \in B_n$ as above, the inequalities (21) imply

$$\begin{aligned}&\left|d\widehat{\mu}^{\mathsf{Vovk}}_{n,\eta}(\theta) - d\pi_n(\theta)\right|/d\theta = \exp\left(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n)\right)\left|\frac{\exp(-nr_n(\theta))}{Z^{\mathsf{Vovk}}_{\Theta,n}} - \frac{1}{Z_n^{\mathsf{N}}}\right| \\ &\le \frac{\exp(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n))}{Z_n^{\mathsf{N}}}\left[\left|\exp(-nPM_{\mathrm{Lip},2}\cdot\delta_n^3)\frac{Z^{\mathsf{N}}_{\Theta,n}}{Z_n^{\mathsf{Vovk}}} - 1\right| + \left|\frac{1}{Z_n^{\mathsf{N}}} - \frac{1}{Z^{\mathsf{N}}_{\Theta,n}}\right|\right]\end{aligned}$$

Integrating over B_n and invoking inequality (21) then implies

$$T_1 = \int_{B_n}|d\widehat{\mu}^{\mathsf{Vovk}}_{n,\eta} - d\pi_n| \le \frac{\int_{B_n}\exp(-\frac{n}{2}(\theta - \widehat{\theta}_n)^T\nabla^2\mathsf{Risk}_n(\widehat{\theta}_n)(\theta - \widehat{\theta}_n))}{Z_n^{\mathsf{N}}}\cdot o(1) \to 0.$$

Lastly, we note that the final term T_3 in the variation distance (19) satisfies $\pi_n(\Theta^c) \to 0$ as $n \to \infty$ as on event $\mathcal{E}_n$, there is eventually a ball of some (fixed) radius $\epsilon > 0$ such that $\widehat{\theta}_n + \epsilon \mathbb{B}_2^d \subset \Theta$, and $\nabla^2 \mathsf{Risk}_n(\widehat{\theta}_n) \succeq (\lambda_{\min}/2) I$. For Standard normal concentration results then immediately imply that $\pi_n(\Theta^c) \le \pi_n(\{\widehat{\theta}_n + \epsilon \mathbb{B}_2^d\}) \to 0$, as the variance of $\theta \sim \pi_n$ satisfies $\mathbb{E}_{\pi_n}[\|\theta - \mathbb{E}[\theta]\|_2^2] \le C/n$ for some problem-dependent C. We conclude that each of $T_1, T_2, T_3 \to 0$ in the variation distance (19).

B.4 Proof of Corollary 7

We again use the event $\mathcal{E}_n$ in Eq. (18) in the proof of Theorem 6 and $\frac{\log\log n}{n} \ll \delta_n^2 \ll \frac{1}{\sqrt{n}}$ as well. Let $p_n = \widehat{P}_{n,\eta}^{\mathsf{Vovk}}$ and $\mu_n = \widehat{\mu}_{n,\eta}^{\mathsf{Vovk}}$ for shorthand, and let $p_{\widehat{\theta}_n}$ the the point model. Let $B_n = \widehat{\theta}_n + n^{-1/4}\mathbb{B}_2^d$ be a ball of radius $n^{-1/4}$ around $\widehat{\theta}_n$, where for all large enough n, on $\mathcal{E}_n$ we have $B_n \subset B \subset \Theta$, where we recall that B is the neighborhood of $\theta^\star$ in Assumption 1. Then for the base measure ν on $\mathcal{Y}$, we expand

$$\begin{aligned} 2\left\|p_n(\cdot \mid x) - p_{\widehat{\theta}_n}(\cdot \mid x)\right\|_{\mathrm{TV}} &= \int \left| \int_\Theta \left(p_\theta(y \mid x) - p_{\widehat{\theta}_n}(y \mid x)\right) d\mu_n(\theta) \right| d\nu(y) \\ &\le \mu_n(\Theta \setminus B_n) + \int \left| \int_{B_n} \left(p_\theta(y \mid x) - p_{\widehat{\theta}_n}(y \mid x)\right) d\mu_n(\theta) \right| d\nu(y). \end{aligned}$$

By Theorem 6, we have $\mu_n(\Theta \setminus B_n) \to 0$ on $\mathcal{E}_n$. Now, let $\ell_\theta = \log p_\theta$ for shorthand, and also define the shorthands $\dot{p}_\theta = \nabla_\theta p_\theta$ and $\dot{\ell}_\theta = \nabla_\theta \ell_\theta = \frac{\dot{p}_\theta}{p_\theta}$. The Lipschitz condition on $\log p_\theta$ in Assumption 1 guarantees that (for large n) on the set B_n we have $|\frac{\dot{p}_\theta(y|x)}{p_\theta(y|x)}| \le \mathrm{Lip}_p(x, y)$ for $\theta \in B_n$. Writing

$$p_\theta(y \mid x) - p_{\widehat{\theta}_n}(y \mid x) = \int_0^1 \dot{p}_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x)^T (\theta - \widehat{\theta}_n) dt = \int_0^1 \dot{\ell}_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x)^T (\theta - \widehat{\theta}_n) p_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x) dt,$$

we have

$$|p_\theta(y \mid x) - p_{\widehat{\theta}_n}(y \mid x)| \le \mathrm{Lip}_p(x, y) \|\theta - \widehat{\theta}_n\|_2 \int_0^1 p_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x) dt.$$

Thus

$$\begin{aligned} &\int_{\mathcal{Y}} \left| \int_{B_n} \left(p_\theta(y \mid x) - p_{\widehat{\theta}_n}(y \mid x)\right) d\mu_n(\theta) \right| d\nu(y) \\ &\le \int_{\mathcal{Y}} \int_{B_n} \mathrm{Lip}_p(x, y) \|\theta - \widehat{\theta}_n\|_2 \int_0^1 p_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x) dt d\mu_n(\theta) d\nu(y) \\ &= \int_0^1 \int_{B_n} \|\theta - \widehat{\theta}_n\|_2 \left[\int_{\mathcal{Y}} \mathrm{Lip}_p(x, y) p_{t\theta + (1-t)\widehat{\theta}_n}(y \mid x) d\nu(y) \right] d\mu_n(\theta) dt \le \mathrm{Lip}_p(x) n^{-1/4} \end{aligned}$$

on $\mathcal{E}_n$, and we have the desired convergence.