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. Duchi, 2018, 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*

$$\mathsf{sep}(f_1, f_2, \Theta) := \sup \left\{ \varepsilon \ge 0 \mid \begin{array}{l} f_1(\theta) \le \inf_{\theta \in \Theta} f_1(\theta) + \varepsilon \ \ \text{implies} \ \ f_2(\theta) > \inf_{\theta\in\Theta} f_2(\theta) + \varepsilon \\ f_2(\theta) \le \inf_{\theta \in \Theta} f_2(\theta) + \varepsilon \ \ \text{implies} \ \ f_1(\theta) > \inf_{\theta\in\Theta} f_1(\theta) + \varepsilon, \end{array} \text{all } \theta \in \Theta \right\}.$$

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 (Duchi, 2018, 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*

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

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

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 Duchi (2018); Wainwright (2019), is to show that for some constants $0 < c_0, c_1 < \infty$ and a power $\beta \ge 0$, we can choose $P_{\pm 1}$ to scale with a desired rate ε via

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

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 $\delta \ge 0$ jointly satisfy*

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

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

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

while

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

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 $Q_{\pm\gamma} := (1 - \gamma) P_0 + \gamma P_{\pm 1}$,*

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

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

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

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

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

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

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

where the remainder term $|\operatorname{rem}(\delta)| \le \sup_{|\Delta|\le\delta} \ell''(t+\Delta, y)$. 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

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

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

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

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

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

and therefore $\mathsf{Risk}^{\Theta}_{Q_{\pm\gamma}}(\theta) = \gamma \mathsf{Risk}^{\Theta}_{P_{\pm 1}}(\theta)$. 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 $q^\star_\ell \equiv q^\star_\ell(t,y) > 0$, so that we obtain the lower bound

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

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 $|\ell''(t+\Delta, y)| \ge |\ell'''(t,y)||\Delta|/2$, and so in particular for all small δ,

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

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

B Technical appendices

B.1 Derivations for Example 2

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 $\text{sign}(\ell'(t,y)) = 1 \ne -1 = \text{sign}(\ell'(\alpha t, y_0))$. 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

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

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

$$\mathfrak{M}_n(\Theta, \Gamma, \gamma) \ge c \min\left\{\frac{\sqrt{\gamma RB}}{\sqrt{n}}, \frac{\exp(2RB/5)}{n}\right\}. \tag{16}$$

B.2 Derivations for Example 3

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 $b \ge 1$ and that $\operatorname{diam}(\mathcal{Y}) := \max\{y \in \mathcal{Y}\}$ is finite and at least 3.

First, we construct a lower bound on $q_\ell^\star(t, y)$: 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

$$q_\ell^\star(t, y) \ge \frac{-\frac{y+1}{b}\frac{1}{1+e} + \frac{1}{b}}{-\frac{y+1}{b}\frac{1}{1+e} + \frac{1}{b} - (y + 1)\frac{e^b}{1+e^b} + 1} = \frac{\frac{y+1}{1+e} - 1}{\frac{y+1}{1+e} - b + 1 + b(y + 1)\frac{e^b}{1+e^b}} \gtrsim \frac{1}{b}$$

as $b \ge 1$ and $y \ge 3$. Additionally, we have $|\ell'(t, y)| \gtrsim y$ and $\ell''(t, y) \lesssim ye^{-b}$, and so, as in the derivation in Example 2 and by setting $\delta \gtrsim b$, we obtain that there exist numerical constants $c_0, c_1 > 0$ such that

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

Substituting, we obtain the analogue of inequality (16), that is,

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

B.3 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

$$\frac{1}{2p^{3/2}} \ge \eta(1 - 1/\sqrt{p})^2 \quad \text{or} \quad \frac{1}{2} \ge \eta(p^{3/2} - 2p + \sqrt{p})$$

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 $1 \ge 2\eta(1/27 - 2/9 + 1/3) = \frac{8}{27}\eta$, or $\eta \le \frac{27}{8}$.

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 $\eta \le \frac{1}{2}$.

B.4 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 $(X_i,Y_i)\overset{\text{iid}}{\sim}P_v$ 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. (Duchi, 2018, Sec. 5) or (Wainwright, 2019, Ch. 15)).

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

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

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, $\widehat{\Psi}\ne V$, for which we use Fano's inequality Cover and Thomas (2006):

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 $(X_i,Y_i)\overset{\text{iid}}{\sim}P_v$ conditional on $V=v$ and V is uniform on $\mathcal{V}$. Then for any $\widehat{\Psi}$,*

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

Now, we define the collection of problems we consider and their induced risks. Let X be uniform on $\{\pm1\}^d$, and let

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

be the density of P_θ with respect to the base measure ν. For a value $\delta\ge0$ 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.4.1 and B.4.2.

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 $\delta\ge0$ and $u,v\in\mathbb{R}^d$,*

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

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 $p_{\delta v}(y\mid x)=\exp(y\theta^Tx-A(\theta^Tx))$. 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 $\left\|u\right\|_2\le1$, $\left\|v\right\|_2\le1$,*

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

With these two lemmas, the result is relatively straightforward. We consider two cases: that $d\ge8$ and (for completeness) that $d\le8$, which we defer temporarily. Let $d\ge8$. By a standard volume argument (Wainwright, 2019, 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 $|\mathcal{V}|\ge\exp(d/4)$ and $\left\|v-w\right\|_2\ge\frac{1}{2}$ for each $v\ne w\in\mathcal{V}$. 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

$$I(V;X_1^n,Y_1^n)\le\frac{1}{|\mathcal{V}|^2}\sum_{v,w\in\mathcal{V}}D_{\rm kl}\left(P_v^n\|P_w^n\right)\overset{(\star)}{\le}n\cdot\frac{1}{|\mathcal{V}|^2}\sum_{v,w\in\mathcal{V}}\delta^2A''(0)\left\|v-w\right\|_2^2\le4n\delta^2A''(0),$$

where inequality $(\star)$ holds for any sufficiently small $\delta\ge0$ by Lemma 14. Applying Lemmas 11 and 13 by noting that $\|v - w\|_2 \ge \frac{1}{2}$, there exists a numerical constant $c > 0$ such that for small enough $\delta \ge 0$,

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

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.4.1 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

$$\begin{aligned}2\,\mathsf{sep}(\mathsf{Risk}_{\delta v}, \mathsf{Risk}_{\delta w}, \mathcal{P}) &\ge \mathbb{E}\left[D_{\rm kl}\left(p_{\delta v}(\cdot \mid X) \| (1/2)(p_{\delta v}(\cdot \mid X) + p_{\delta w}(\cdot \mid X))\right)\right] \\ &\quad + \mathbb{E}\left[D_{\rm kl}\left(p_{\delta w}(\cdot \mid X) \| (1/2)(p_{\delta v}(\cdot \mid X) + p_{\delta w}(\cdot \mid X))\right)\right],\end{aligned} \tag{17}$$

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 (17). 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

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

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

$$D_{\rm kl}\left(p_t \| (1/2)(p_t + p_u)\right) = \frac{1}{8}(u-t)^2 A''(0) \pm C(A)|u-t|^3, \tag{18}$$

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 (17) and using that $\mathbb{E}[XX^T] = I_d$ by construction then gives the lemma.

B.4.2 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.5 Proof of Theorem 6

Recall $\widehat{\theta}_n = \mathrm{argmin}_{\theta \in \Theta} \mathsf{Risk}_n(\theta)$ 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 $M_{\mathrm{Lip},2}(x,y)$ Lipschitz (Def. 3.1), and for an $\epsilon > 0$ to be determined

$$\begin{aligned}
&\mathcal{E}_n := \\
&\left\{ P_n M_{\mathrm{Lip},2} \le 2 P M_{\mathrm{Lip},2},\ \frac{\lambda_{\min}}{2} I \preceq \nabla^2 \mathsf{Risk}_n(\theta) \preceq 2\lambda_{\max} I \text{ for } \|\theta - \theta^\star\|_2 \le \epsilon,\ \|\widehat{\theta}_n - \theta^\star\|_2 \le \delta_n \right\}.
\end{aligned} \tag{19}$$

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

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

Finally, we argue that $\|\widehat{\theta}_n - \theta^\star\|_2 \le \delta_n$ eventually. A standard argument (van der Vaart, 1998, Thm. 5.7) and the Glivenko Cantelli theorem, which implies $\sup_{\theta \in \Theta} |\mathsf{Risk}_n(\theta) - \mathsf{Risk}_P(\theta)| \overset{a.s.}{\to} 0$ by the compactness of Θ, gives the consistency $\widehat{\theta}_n \overset{a.s.}{\to} \theta^\star$. 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

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

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

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

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

$$2\left\|\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - \pi_n\right\|_{\mathrm{TV}} = \underbrace{\int_{\widehat{\theta}_n + B_n} |d\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - d\pi_n|}_{=:T_1} + \underbrace{\int_{\Theta \setminus \{\widehat{\theta}_n + B_n\}} |d\widehat{\mu}_{n,\eta}^{\mathsf{Vovk}} - d\pi_n|}_{=:T_2} + \underbrace{\pi_n(\Theta^c)}_{=:T_3}. \tag{20}$$

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 $\frac{\lambda_{\min}}{4} - 2c^2\lambda_{\max} =: K > 0$,

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

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

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

where the inequality uses the definition (19) 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 (20) by

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

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

Let us turn to term T_1 in expression (20). For sets $A \subset \mathbb{R}^d$ we define the normalizing constants

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

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

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

and

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

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

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

where as before $c_d = \log\operatorname{Vol}(\mathbb{B}_2^d)$, 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

$$\frac{Z^{\mathsf{Vovk}}_{B_n,n}}{Z^{\mathsf{N}}_{B_n,n}} + \rho_n \ge \frac{Z^{\mathsf{Vovk}}_{B_n,n} + Z^{\mathsf{Vovk}}_{\Theta\setminus B_n,n}}{Z^{\mathsf{N}}_{B_n,n}} \ge \frac{Z^{\mathsf{Vovk}}_{\Theta,n}}{Z^{\mathsf{N}}_{\Theta,n}} \ge \frac{Z^{\mathsf{Vovk}}_{B_n,n}}{Z^{\mathsf{N}}_{B_n,n} + Z^{\mathsf{N}}_{\Theta\setminus B_n,n}} \ge \left(\frac{Z^{\mathsf{N}}_{B_n,n}}{Z^{\mathsf{Vovk}}_{B_n,n}} + \rho_n\right)^{-1}. \tag{21}$$

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 (21), we find that for all large enough n, on the event $\mathcal{E}_n$ in Eq. (19), we have the bounds

$$\exp(-PM_{\mathrm{Lip},2}\delta_n^3) - O(1)\rho_n \le \frac{Z^{\mathsf{Vovk}}_{\Theta,n}}{Z^{\mathsf{N}}_{\Theta,n}} \le \exp(PM_{\mathrm{Lip},2}\cdot\delta_n^3) + O(1)\rho_n. \tag{22}$$

Finally, we return to computing the densities in the term T_1 in Eq. (20). Let $Z_n^{\mathsf{N}} = Z_{\mathbb{R}^d,n}^{\mathsf{N}}$, where an argument similar to those above shows that $Z_n^{\mathsf{N}} / Z_{\Theta,n}^{\mathsf{N}} \to 1$ 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 $\|\mathrm{rem}_n(\theta)\|_2 \le P M_{\mathrm{Lip},2} \cdot \delta_n^3$ for any $\theta \in B_n$ as above, the inequalities (22) imply

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

Integrating over B_n and invoking inequality (22) then implies

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

Lastly, we note that the final term T_3 in the variation distance (20) 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 (20).

B.6 Proof of Corollary 7

We again use the event $\mathcal{E}_n$ in Eq. (19) 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.