A The Assouad and Fano Methods for Minimax Lower Bounds

In this precursor to the appendix, we review the Le Cam, Fano and Assouad methods [2, 29, 1, 27] for proving lower bounds for stochastic optimization. Each reduces estimation to testing then uses information theoretic tools to bound the probability of error in various hypothesis tests.

A.1 Le Cam and Fano Methods

We start with a lemma that provides the standard reduction from estimation to testing that we extensively use in our proofs. This is essentially [12, Ex. 7.5]; we provide the proof for completeness.

Lemma 1 (From estimation to testing)**.** *Let $\mathcal{P}$ be a collection of distributions over $\mathcal{X}$ and $L : \Theta \times \mathcal{P} \to \mathbf{R}_+$ satisfy*

$$\inf_{\theta \in \Theta} L(\theta, P) = 0 \text{ for } P \in \mathcal{P}.$$

For distributions $P, Q \in \mathcal{P}$, define the separation

$$\mathsf{sep}_L(P, Q; \Theta) := \sup \left\{ \delta \ge 0 \,\middle|\, \text{for all } \theta \in \Theta, \begin{array}{l} L(\theta, P) \le \delta \text{ implies } L(\theta, Q) \ge \delta \\ L(\theta, Q) \le \delta \text{ implies } L(\theta, P) \ge \delta \end{array} \right\}.$$

Let $\delta > 0$ and $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ be a family of distributions indexed by a finite set $\mathcal{V}$ satisfying the separation condition $\mathsf{sep}_L(P_v, P_{v'}; \Theta) \ge \delta$ for $v \neq v' \in \mathcal{V}$. Then for $X_1^n \stackrel{\text{iid}}{\sim} P$,

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{\theta}(X_1^n), P) \ge \delta \inf_{\psi} \mathbb{P}(\psi(X_1^n) \neq V),$$

where $\mathbb{P}$ is the joint distribution over the random index V chosen uniformly in $\mathcal{V}$ and $X_1^n \stackrel{\text{iid}}{\sim} P_v$ conditional on $V = v$.

Proof. Let $V \sim \mathsf{Uniform}(\mathcal{V})$ and $X_1^n \mid (V = v) \stackrel{\text{iid}}{\sim} P_v$. Then for any estimator $\widehat{\theta}$, we have

$$\sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{\theta}(X_1^n), P) \ge \frac{1}{|\mathcal{V}|} \sum_v \mathbf{E}_{P_v} L(\widehat{\theta}, P_v) \ge \delta \frac{1}{|\mathcal{V}|} \sum_v P_v(L(\widehat{\theta}, P_v) \ge \delta) = \delta \mathbb{P}(L(\widehat{\theta}(X_1^n), P_V) \ge \delta),$$

where $\mathbb{P}$ denotes the joint distribution of X_1^n and V. Define the test $\psi(x_1^n) := \operatorname{argmin}_{v \in \mathcal{V}} L(\widehat{\theta}(x_1^n), P_v)$. The separation assumption guarantees that if $\psi(\theta) \neq v$ then $L(\theta, P_v) \ge \delta$, so

$$\mathbb{P}(L(\widehat{\theta}(X_1^n), P_V) \ge \delta) \ge \mathbb{P}\left(\psi(X_1^n) \neq V\right).$$

Taking the infimum over all tests ψ yields the result. □

With this, the classical Le Cam and Fano methods are straightforward combinations of Lemma 2 with (respectively) Le Cam's lemma [29, Lemma 1] and Fano's inequality [8, Theorem 2.10.1].

Proposition 6 (Le Cam's method)**.** *Let P_0 and P_1 be two distributions of $\mathcal{P}$ over $\mathcal{X}$. Let $\delta > 0$ be such that $\mathsf{sep}_L(P_0, P_1, \Theta) \ge \delta$. Then*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{\theta}(X_1^n), P) \ge \frac{\delta}{2}(1 - \|P_0^n - P_1^n\|_{\mathsf{tv}}).$$

Proposition 7 (Fano's method)**.** *Let $\mathcal{V}$ be a finite index set and $\{P_v\}_{v \in \mathcal{V}}$ a collection of distributions contained by $\mathcal{P}$ such that $\min_{v \neq v'} \mathsf{sep}_L(P_v, P_{v'}, \Theta) \ge \delta$, then*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{\theta}(X_1^n), P) \ge \delta \left(1 - \frac{\mathsf{I}(X_1^n; V) + \log 2}{\log |\mathcal{V}|}\right).$$

With these tools, minimax lower bounds on the stochastic risk $\mathfrak{M}_n^{\mathsf{S}}$ in Section 2 follow by (i) demonstrating an appropriate loss L and (ii) separation. The next lemma, essentially present in the paper [1] (cf. [11]), reduces optimization to testing by providing an appropriate separation function.

Lemma 2 (From optimization to function estimation)**.** *Let $\mathcal{X}$ be a sample space, $\Theta \subset \mathbf{R}^d$, $\mathcal{F}$ be a collection a functions $\mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$, and $\mathcal{P}$ be a collection of distributions over $\mathcal{X}$. Let $\mathcal{V}$ index $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$. For $F \in \mathcal{F}$, define $f_v(\theta) := \mathbf{E}_{P_v}[F(\theta, X)]$ and for each $v, v' \in \mathcal{V}$, set*

$$\mathsf{d}_{\text{opt}}(v, v', \Theta) := \inf_{\theta\in\Theta} \left\{ f_v(\theta) + f_{v'}(\theta) - \inf_{\theta\in\Theta} f_v(\theta) - \inf_{\theta\in\Theta} f_{v'}(\theta) \right\}.$$

If $\mathsf{d}_{\text{opt}}(v, v', \Theta) \geq \delta \geq 0$ for all $v \neq v' \in \mathcal{V}$, then

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}) \geq \mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}, \mathcal{P}) \geq \frac{\delta}{2} \inf_{\psi} \mathbb{P}(\psi(X_1^n) \neq V).$$

Proof. We construct an appropriate loss L and apply Lemma 1. Define $L(\theta, P) := f_P(\theta) - \inf_{\theta\in\Theta} f_P(\theta)$. By construction, $L(\theta, P) \geq 0$ and $\inf_{\theta\in\Theta} L(\theta, P) = 0$ for all $\theta \in \Theta$ and $P \in \mathcal{P}$. Let $v \neq v' \in \mathcal{V}$. Then if $L(\theta, P_v) = f_v(\theta) - \inf_{\theta\in\Theta} f_v(\theta) \leq \frac{1}{2}\mathsf{d}_{\text{opt}}(v, v', \Theta)$, it is evidently the case that $f_{v'}(\theta) - \inf_{\theta\in\Theta} f_{v'}(\theta) \geq \frac{1}{2}\mathsf{d}_{\text{opt}}(v, v', \Theta)$, so that $\mathsf{sep}_L(P_v, P_{v'}, \Theta) \geq \frac{1}{2}\mathsf{d}_{\text{opt}}(v, v', \Theta)$. The distributions $\{P_v\}_{v\in\mathcal{V}}$ are $\delta/2$-separated, allowing application of Lemma 1. □

Our general strategy for proving lower bounds on $\mathfrak{M}_n^{\mathsf{S}}$ is as follows:

- Choose a function $F \in \mathcal{F}$ and define $\mathcal{V}$ and $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ such that $\mathsf{d}_{\text{opt}}(v, v', \Theta) \geq \delta > 0$.
- Lower bound the testing error $\inf_\psi \mathbb{P}(\psi(X_1^n) \neq V)$, and choose the largest separation δ to make this testing error a positive constant.

To showcase this proof technique, we prove that minimax stochastic risk for 1-dimensional optimization has lower bound $1/\sqrt{n}$; we use this to address technicalities in later proofs.

Lemma 3. *Let $\mathcal{F}^{d=1} = \{f : \mathbf{R} \times \mathcal{X} \to \mathbf{R} \mid f(\cdot, x)$ is convex and 1-Lipschitz$\}$. Then*

$$\mathfrak{M}_n^{\mathsf{S}}([-1, 1], \mathcal{F}^{d=1}) \geq \frac{1}{4\sqrt{6n}}.$$

Proof. Let $\Theta = [-1, 1]$ and $\mathcal{X} = \{\pm 1\}, \mathcal{V} = \{\pm 1\}$.

To see the separation condition, let $F(\theta, x) := |\theta - x|$. For $\delta \in [0, \frac{1}{2}]$, we define P_v s.t. if $X \sim P_v$ we have

$$X = \begin{cases} 1 & \text{with probability } \frac{1+v\delta}{2} \\ -1 & \text{with probability } \frac{1-v\delta}{2}. \end{cases}$$

We have $f_v(\theta) = \frac{1+\delta}{2}|\theta - v| + \frac{1-\delta}{2}|\theta + v|$ and $\inf_\theta f_v(\theta) = \frac{1-\delta}{2}$. To lower bound the separation, note that

$$f_1(\theta) + f_{-1}(\theta) - \inf_{\Theta} f_1 - \inf_{\Theta} f_{-1} = |\theta - 1| + |\theta + 1| - (1 - \delta) \geq \delta.$$

This yields $\mathsf{d}_{\text{opt}}(1, -1, \Theta) \geq \delta$.

We lower bound the testing error via Proposition 6:

$$\inf_{\psi:\mathcal{X}^n\to\{\pm 1\}} \mathbb{P}(\psi(X_1^n) \neq V) = \frac{1}{2}(1 - \left\| P_1^n - P_{-1}^n \right\|_{\mathsf{tv}}) \geq \frac{1}{2}\left(1 - \sqrt{\frac{n}{2} D_{\text{kl}}\left(P_1 \| P_{-1}\right)}\right),$$

where the rightmost inequality is Pinsker's inequality. Noting that $D_{\text{kl}}\left(P_1 \| P_{-1}\right) = \delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, \frac{1}{2}]$ and setting $\delta = 1/\sqrt{6n}$ yields the result. □

A.2 The Assouad Method

Assouad's method reduces the problem of estimation (or optimization) to one of multiple binary hypothesis tests. In this case, we index a set of distributions $\mathcal{P} = \{P_v\}_{v\in\mathcal{V}}$ on a set $\mathcal{X}$ by the hypercube $\mathcal{V} = \{\pm 1\}^d$. For a function $F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$, we define $f_v(\theta) := \mathbf{E}_{P_v}[F(\theta, X)]$. Then for a vector $\delta \in \mathbf{R}_+^d$, following Duchi [11, Lemma 5.3.2], we say that the functions $\{f_v\}$ induce a δ-separation in Hamming metric if

$$f_v(\theta) - \inf_{\theta\in\Theta} f_v(\theta) \geq \sum_{j=1}^{d} \delta_j 1(\text{sign}(\theta_j) \neq v_j)\,. \tag{8}$$

With this condition, we have the following generalized Assouad method [11, Lemma 5.3.2].

Lemma 4 (Generalized Assouad's method)**.** *Let $X_1^n \overset{\text{iid}}{\sim} P_V$, where $V \sim \mathsf{Uniform}(\{\pm 1\}^d)$. Define the averages*

$$\mathbb{P}_{+j} := \frac{1}{2^{d-1}} \sum_{v : v_j = 1} P_v^n \;\; \textit{and} \;\; \mathbb{P}_{-j} := \frac{1}{2^{d-1}} \sum_{v : v_j = -1} P_v^n.$$

Assume that the collection $\{f_v\}$ for $f_v = \mathbf{E}_{P_v}[F(\cdot, X)]$ induces a δ-separation (8). Then letting $\mathcal{F} = \{F\}$, the single function F,

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}, \mathcal{P}) \geq \frac{1}{2} \sum_{j=1}^d \delta_j (1 - \|\mathbb{P}_{+j} - \mathbb{P}_{-j}\|_{\mathsf{tv}}).$$

B Proofs for Section 3.1

B.1 Proof of Proposition 1

We use the general information-theoretic framework of reduction from estimation to testing presented in Section A.1 to prove the lower bound.

Separation Let us consider the sample space $\mathcal{X} = \{\pm e_j\}_{j \leq d}$ and the function $F(\theta, x) := \theta^\top x$; F belongs to $\mathcal{F}^{\gamma,1}$. Let $\delta \in [0, 1/2]$, for $v \in \{\pm 1\}^d$, we define P_v such that for $X \sim P_v$ we have

$$X = \begin{cases} v_j e_j & \text{with probability } \frac{1+\delta}{2d} \\ -v_j e_j & \text{with probability } \frac{1-\delta}{2d}. \end{cases}$$

We then have $f_v(\theta) = \frac{\delta}{d} \theta^\top v$. By duality,

$$f_v^* := \inf_\Theta f_v = -\frac{\delta}{d} \sup_{\theta \in \mathbf{B}_p(0,1)} v^\top \theta = -\frac{\delta}{d} \|v\|_{p^*},$$

where p^* is such that $1/p + 1/p^* = 1$. For $v, v' \in \{\pm 1\}^d$, we thus have:

$$\begin{aligned} \mathsf{d}_{\text{opt}}(v, v', \Theta) = \inf_{\theta \in \Theta} f_v(\theta) + f_{v'}(\theta) - f_v^* - f_{v'}^* &= \inf_{\theta \in \mathbf{B}_p(0,1)} \frac{\delta}{d} (\theta^\top (v + v') + \|v\|_{p^*} + \|v'\|_{p^*}) \\ &= \frac{\delta}{d} (\|v\|_{p^*} + \|v'\|_{p^*} - \|v + v'\|_{p^*}) \\ &= 2\frac{\delta}{d} \left[d^{1/p^*} - (d - \mathrm{d}_{\mathrm{Ham}}(v, v'))^{1/p^*} \right], \end{aligned}$$

where $\mathrm{d}_{\mathrm{Ham}}(v, v')$ is the Hamming distance between v and v'. The Gilbert-Varshimov bound [12, Lemma 7.5] guarantees the existence of a $d/2$ ℓ_1-packing of $\{\pm 1\}^d$ of size at least $\exp(d/8)$. Let $\mathcal{V}$ be such a packing; we have that, for a numerical constant $c_0 > 0$:

$$\forall v \neq v' \in \mathcal{V}, \mathsf{d}_{\text{opt}}(v, v', \Theta) \geq c_0 \delta d^{-1/p}. \tag{9}$$

Applying Lemma 2 yields

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \frac{c_0}{2} \delta d^{-1/p} \inf_\psi \mathbb{P}(\psi(X_1^n) \neq V).$$

Bounding the testing error We bound the testing error with Fano's inequality and upper bounding the mutual information $\mathsf{I}(X; V)$. Using the identity $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$, it holds

$$\mathsf{I}(X_1^n; V) \leq n \max_{v, v'} D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) \leq 3n\delta^2,$$

and, recalling that $\log |\mathcal{V}| \geq d/8$ yields

$$\inf_\psi \mathbb{P}(\psi(X_1^n) \neq V) \geq \left(1 - \frac{3n\delta^2 + \log 2}{d/8}\right).$$

In the case that $d \geq 32 \log 2$, choosing $\delta = \sqrt{\frac{d}{48n}}$ yields the desired lower-bound. In the case that $d < 32 \log 2$, with $\mathcal{F}^{d=1}$ as in Lemma 3, that any 1-dimensional optimization problem may be embedded into a d-dimensional problem yields

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \mathfrak{M}_n^{\mathsf{S}}([-1,1], \mathcal{F}^{d=1}) \gtrsim \frac{1}{\sqrt{n}}.$$

This gives the lower bound for all $d \in \mathbf{N}$.

To conclude the proof, we establish an upper bound on the minimax regret. We consider the regret guarantee of (5) for $h(\theta) = \frac{1}{2}\|\theta\|_2^2$. Since $p \geq 2$, it holds that for all $\theta \in \mathbf{R}^d$, $\|\theta\|_2 \leq d^{\frac{1}{2}-\frac{1}{p}}\|\theta\|_p$ and thus $\sup_{\theta,\theta' \in \Theta} \mathrm{D}_h(\theta, \theta') \leq d^{\frac{1}{2}-\frac{1}{p}}$. On the other hand, since $r \in [1,2]$, $\|g\|_2 \leq \|g\|_r \leq 1$. A straightforward optimization of the stepsize α yields the upper bound on $\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma)$. □

B.2 Proof of Proposition 2

The proof is very similar to Proposition 1 so we forego some of the details.

Separation We consider $\mathcal{X} = \{\pm 1\}^d$ and $F(\theta, x) := \eta\theta^\top x$—we will decide the value of η later in the proof. For $v \in \{\pm 1\}^d$, we define P_v such that for $X \sim P_v$ we have

$$X_j = \begin{cases} v_j & \text{with probability} \frac{1+\delta}{2} \\ -v_j & \text{with probability} \frac{1-\delta}{2}. \end{cases}$$

This yields $f_v(\theta) = \eta\delta\theta^\top v$. Considering again the Gilbert-Varshimov packing $\mathcal{V} \subset \{\pm 1\}^d$, we lower bound the separation

$$\text{for all } \ v \neq v' \in \mathcal{V}, \mathsf{d}_{\text{opt}}(v, v', \Theta) = \inf_{\theta \in \Theta} f_v(\theta) + f_{v'}(\theta) - f_v^* - f_{v'}^* \geq c_0 \eta\delta d^{1/p^*}.$$

Bounding the testing error Noting that

$$D_{\text{kl}}\left(P_v \| P_{v'}\right) = \sum_{j \leq d} \mathbf{1}_{v_j = v_j'} \delta \log \frac{1+\delta}{1-\delta} \leq 3d\delta^2,$$

and have $\mathsf{I}(X_1^n; V) \leq 3nd\delta^2$. For F to remain in $\mathcal{F}^{\gamma,1}$, we must have that for all $x \in \mathcal{X}, \eta\|x\|_r \leq 1$; noting that $\|x\|_r = d^{1/q}$, we choose $\eta = d^{-1/q}$. In the case that $d \geq 32 \log 2$, choosing $\delta = 1/\sqrt{48n}$ yields the minimax lower-bound

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \gtrsim \frac{d^{\frac{1}{p^*}} d^{-\frac{1}{q}}}{\sqrt{n}} = \frac{d^{\frac{1}{2}-\frac{1}{p}} d^{\frac{1}{2}-\frac{1}{q}}}{\sqrt{n}}.$$

In the case that $d < 32 \log 2$, we once again refer Lemma 3, which concludes the proof for the lower bound on the minimax stochastic risk.

For the upper bound, we turn to (5), with $h(\theta) = \frac{1}{2}\|\theta\|_2^2$. It holds again that $\sup_{\theta,\theta' \in \Theta} \mathrm{D}_h(\theta, \theta') \leq d^{1/2-1/p}$. Since $r \geq 2$, we have that $\sup_{\|g\|_r \leq 1} \|g\|_2 = d^{\frac{1}{2}-\frac{1}{r}}$ and choosing the stepsize α to optimize (5) yields the upper bound on the minimax regret. □

C Proofs for Section 3.2

C.1 Proof of Theorem 1

For the upper bound, we use Corollary 1. Because $\mathbf{B}_\gamma(0,1)$ is quadratically convex, we have $\mathsf{QHull}(\mathbf{B}_\gamma(0,1)) = \mathbf{B}_\gamma(0,1)$, so that $\sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1))} \theta^\top g = \gamma^*(\theta)$, giving the upper bound. The lower bound uses Proposition 3. Define the hyperrectangle $\mathsf{Rec}(\theta) := \prod_{j \leq d}[-|\theta_j|, |\theta_j|]$, so that, by orthosymmetry of Θ, $\Theta \supset \mathsf{Rec}(\theta)$ for all $\theta \in \Theta$. Additionally, recalling the notation (3) of $\mathcal{F}^{\gamma,1}$ and $\mathcal{F}^M$, if $M \in \mathbf{R}_+^d$ satisfies $\gamma(M) \leq 1$ then, by orthosymmetry of γ, $\mathcal{F}^{\gamma,1} \supset \mathcal{F}^M$. Thus

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \mathfrak{M}_n^{\mathsf{S}}(\mathsf{Rec}(\theta), \gamma) \geq \mathfrak{M}_n^{\mathsf{S}}(\mathsf{Rec}(\theta), \mathcal{F}^M) \geq \frac{1}{8\sqrt{n \log 3}} \sum_{j \leq d} |\theta_j| M_j$$

for all $M \in \mathbf{B}_\gamma(0,1) \cap \mathbf{R}^d_+$ and $\theta \in \Theta$. Taking a supremum over $M \in \mathbf{B}_\gamma(0,1)$ and $\theta \in \Theta$, we have

$$\mathfrak{M}^{\mathsf{S}}_n(\Theta,\gamma) \geq \frac{1}{8\sqrt{n\log 3}} \sup_{\theta\in\Theta} \sup_{\gamma(M)\leq 1} \theta^\top M = \frac{1}{8\sqrt{n\log 3}} \sup_{\theta\in\Theta} \gamma^*(\theta).$$

□

C.2 Proof of Theorem 2

The upper bound is simply Corollary 1. For the lower bound, similar to our warm-up in Section 3.1, we consider "sparse" gradients, though instead of using Fano's method we use Assouad's method to more carefully relate the geometry of the norm γ and constraint set Θ.

Let a be such that $\mathsf{Rec}(a) \subset \Theta$. We consider the sample space $\mathcal{X} := \{\pm e_j\}_{j\leq d}$ and functions

$$F(\theta,x) := \sum_{j\leq d} \frac{1}{\gamma(e_j)} |x_j||\theta_j - a_j x_j|.$$

For any $x \in \mathcal{X}$, the subdifferential $\partial_\theta F(\theta,x)$ has at most one non-zero coordinate; the orthosymmetry of γ implies $F \in \mathcal{F}^{\gamma,1}$. Let $p \in \mathbf{R}^d_+$ (to be specified presently) be such that $\mathbf{1}^\top p = 1$ and for $1 \leq j \leq d$, let $\delta_j \in [0,1/2]$. We define the distributions P_v on $\mathcal{X}$ by

$$X = \begin{cases} v_j e_j & \text{with probability } \frac{p_j(1+\delta_j)}{2} \\ -v_j e_j & \text{with probability } \frac{p_j(1-\delta_j)}{2}. \end{cases}$$

With this choice, we evidently have

$$f_v(\theta) = \mathbf{E}_{X\sim P_v} F(\theta,X) = \sum_{j\leq d} \frac{p_j}{\gamma(e_j)} \left[\frac{1+\delta_j}{2}|\theta_j - a_j v_j| + \frac{1-\delta_j}{2}|\theta_j + a_j v_j|\right]$$

and immediately that $\inf_\Theta f_v = \sum_{j\leq d} \frac{p_j a_j}{\gamma(e_j)}(1-\delta_j)$. As a consequence, we have the Hamming separation (recall Eq. (8))

$$f_v(\theta) - \inf_\Theta f_v = \sum_{j\leq d} \frac{p_j a_j \delta_j}{\gamma(e_j)} \mathbf{1}_{\mathrm{sign}(\theta_j)\neq v_j},$$

which allows us to apply Assouad's method via Lemma 4.

Using the same notation as Lemma 4, we have

$$\left\|\mathbb{P}^n_{+j} - \mathbb{P}^n_{-j}\right\|^2_{\mathsf{tv}} \leq \frac{1}{2} D_{\mathrm{kl}}\left(\mathbb{P}^n_{+j} \| \mathbb{P}^n_{-j}\right) \leq \log 3 \cdot n p_j \delta_j^2.$$

Choosing $\delta_j = \min\{\frac{1}{2}, \frac{1}{2\sqrt{np_j\log(3)}}\}$ yields the lower bound

$$\mathfrak{M}^{\mathsf{S}}_n(\Theta,\gamma) \geq \frac{1}{8} \sum_{j\leq d} \frac{a_j}{\gamma(e_j)} \min\left\{p_j, \frac{\sqrt{p_j}}{\sqrt{n\log 3}}\right\},$$

and by taking $p_j = (\frac{a_j}{\gamma(e_j)})^2 / \|a/\gamma(e.)\|_2^2$, we obtain for any $a \in \Theta$ that

$$\begin{aligned}
\mathfrak{M}^{\mathsf{S}}_n(\Theta,\gamma) \geq \mathfrak{M}^{\mathsf{S}}_n(\mathsf{Rec}(a),\gamma) &\geq \frac{1}{8} \sum_{j\leq d} \frac{a_j}{\gamma(e_j)} \min\left\{\frac{a_j^2}{\gamma(e_j)^2 \|a/\gamma(e.)\|_2^2}, \frac{1}{\sqrt{n\log 3}} \frac{a_j}{\gamma(e_j)\|a/\gamma(e.)\|_2}\right\} \\
&= \frac{1}{8\left\|a/\gamma(e.)\right\|_2^2} \sum_{j=1}^d \frac{a_j^2}{\gamma(e_j)^2} \min\left\{\frac{a_j}{\gamma(e_j)}, \frac{\|a/\gamma(e.)\|_2}{\sqrt{n\log 3}}\right\}.
\end{aligned}$$

For notational simplicity, define the set $T := \{\theta/\gamma(e.) \mid \theta \in \Theta\}$, which is evidently orthosymmetric and convex (it is a diagonal scaling of Θ). Then

$$\mathfrak{M}^{\mathsf{S}}_n(\Theta,\gamma) \geq \sup_{u\in T} \frac{1}{8\left\|u\right\|_2^2} \sum_{j=1}^d u_j^2 \min\left\{u_j, \frac{\|u\|_2}{\sqrt{n\log 3}}\right\}. \tag{10}$$

For any vector $u \in \mathbf{R}_+^d$ and $c < 1$, if we define $J = \{j \in [d] \mid u_j \geq \frac{c}{\sqrt{d}} \|u\|_2\}$, then

$$\|u\|_2^2 = \|u_J\|_2^2 + \|u_{J^c}\|_2^2 \leq \|u_J\|_2^2 + \|u\|_2^2 \sum_{j \in J^c} \frac{c^2}{d} \leq \|u_J\|_2^2 + c^2 \|u\|_2^2, \text{ i.e. } \|u_J\|_2 \geq \sqrt{1 - c^2} \|u\|_2 .$$

Now, fix $k \in \mathbf{N}$. If in the supremum (10) we consider any vector $u \in T, u \geq 0$ satisfying $\|u\|_0 \leq k$, then setting the index set $J = \{j : u_j \geq \|u\|_2 / \sqrt{n \log 3}\} = \{j : u_j \geq \|u\|_2 / \sqrt{k(n/k) \log 3}\}$ we have

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \frac{1}{8 \|u\|_2^2} \sum_{j=1}^d u_j^2 \min\left\{u_j, \frac{\|u\|_2}{\sqrt{n \log 3}}\right\} \geq \frac{1}{8 \|u\|_2^2} \sum_{j \in J} u_j^2 \frac{\|u\|_2}{\sqrt{n \log 3}} \geq \frac{1}{8}\left(1 - \frac{k}{n \log 3}\right) \frac{\|u\|_2}{\sqrt{n \log 3}}.$$

Taking a supremum over u with $\|u\|_0 \leq k$ gives the theorem.

C.3 Proof of Corollary 2

Given proof of Theorem 2, the proof is nearly immediate. Let $p \in [1, 2], \beta \in (\mathbf{R}_+ \setminus \{0\})^d$ and $\gamma(v) = \|\beta \odot v\|_p$. For the lower bound, the final display of the proof of Theorem 2 above guarantees the lower bound $\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \frac{1}{16} \|u\|_2 / \sqrt{n}$ for all $u \in \{\theta/\gamma(e.) \mid \theta \in \Theta\}$ and $n \geq 2d$. We first observe that $\mathsf{QHull}\left(\mathbf{B}_\gamma(0, 1)\right) = \{v, \|\beta \odot v\|_2 \leq 1\}$. Thus, the upper bound in Theorem 2 is

$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \leq \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \sup_{g : \|\beta \odot g\|_2 \leq 1} \theta^\top g.$$

Using

$$\sup_{g : \|\beta \odot g\|_2 \leq 1} u^\top g = \sup_{z : \|z\|_2 \leq 1} u^\top (z/\beta) = \|u/\beta\|_2 ,$$

and recalling $\beta_j = \gamma(e_j)$ concludes the proof. □

C.4 Proof of Corollary 3

There is a bijective mapping between $\mathcal{F}$ and $\mathcal{F}^{\gamma,1}$: for $F \in \mathcal{F}$, $\theta_0 \in \Theta_0$, and $x \in \mathcal{X}$, we define $\widetilde{F}(\theta_0, x) := F(U\theta_0, x)$. $\operatorname{dom} \widetilde{F} \supset \Theta_0$ and its subdifferential is [16, Thm. 4.2.1]

$$\partial_\theta \widetilde{F}(\theta_0, x) = U^\top \partial_\theta F(U\theta_0, x).$$

Since $\widetilde{F}$ falls within the scope of Theorems 1 or Corollary 2, there exists a diagonal re-scaling Λ^* that achieves the optimal rate. We conclude the proof by observing that a diagonally re-scaled stochastic gradient update on $\widetilde{F}$ corresponds to the update $\theta_{i+1} = \theta_i - U\Lambda^* U^\top g_i$ where $g_i \in \partial_\theta F(\theta_i, X_i)$.

D Proofs for Section 4

D.1 Proof of Theorem 3

Let us tackle the first case stated in the theorem; we reduce the second case to the first one by scaling the dimension.

D.1.1 Case $1 \leq p \leq 1 + 1/\log(2d)$

We always have the lower bound $1/\sqrt{n}$ by Lemma 3 by reducing to a lower-dimensional problem, so we assume without loss of generality that $d \geq 8$.

Separation Let us consider $\mathcal{V} = \{\pm e_j\}_{j \leq d}$. For $v = \pm e_j \in \mathcal{V}$, we define P_v on $X \in \{\pm 1\}^d$ by choosing coordinates of X independently via

$$X_j = \begin{cases} 1 & \text{with probability } \frac{1+\delta v_j}{2} \\ -1 & \text{with probability } \frac{1-\delta v_j}{2}. \end{cases}$$

Immediately, we have $\mathbf{E}_{P_v} X = \delta v$. For $x \in \{\pm 1\}^d$, we define $F(\theta, x) := d^{-1/p^*} \theta^\top x$, so $F \in \mathcal{F}^{\gamma,1}$, $f_v(\theta) = \mathbf{E}_{P_v} F(\theta, X) = \delta d^{-1/p^*} \theta^\top v$, and a calculation gives that $f_v^* := \inf_\Theta f_v = -\delta d^{-1/p^*}$. For $v \neq v' \in \mathcal{V}$, we have

$$\begin{aligned}
\mathsf{d}_{\text{opt}}(v, v', \Theta) = \inf_{\theta \in \Theta} f_v(\theta) + f_{v'}(\theta) - f_v^* - f_{v'}^* &= d^{-1/p^*} \delta \inf_{\theta \in \Theta} \left((v + v')^\top \theta + 2 \right) \\
&= \delta d^{-1/p^*} (2 - \|v + v'\|_{p^*}) \\
&\geq (2 - \sqrt{2}) \delta d^{-1/p^*}.
\end{aligned}$$

Lemma 2 yields

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \frac{2 - \sqrt{2}}{2} \delta d^{-1/p^*} \inf_{\psi : \mathcal{X}^n \to \mathcal{V}} \mathbb{P}(\psi(X_1^n) \neq V).$$

It now remains to bound the testing error.

Bounding the testing error Noting that $|\mathcal{V}| = \log(2d)$, we lower bound the testing error via Fano's inequality

$$\inf_{\psi : \mathcal{X}^n \to \mathcal{V}} \mathbb{P}(\psi(X_1^n) \neq V) \geq \left(1 - \frac{\mathsf{I}(X_1^n; V) + \log 2}{\log(2d)} \right).$$

For any $v \neq v' \in \mathcal{V}$, we have for $\delta \in [0, \frac{1}{2}]$ that

$$D_{\text{kl}}\left(P_v \| P_{v'} \right) = \delta \log \frac{1 + \delta}{1 - \delta} \leq 3\delta^2.$$

We can thus bound the mutual information between X_1^n and V

$$\mathsf{I}(X_1^n; V) \leq n \max_{v \neq v'} D_{\text{kl}}\left(P_v \| P_{v'} \right) \leq 3n\delta^2.$$

In the case that $d < 8$, the lower bound holds trivially via Lemma 3. In the case that $d \geq 8$, assuming that choosing $\delta^2 = \frac{\log(2d)}{6n} \wedge \frac{1}{2}$ yields

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq \frac{2 - \sqrt{2}}{2} d^{-1/p^*} \min \left\{ \sqrt{\frac{\log(2d)}{6n}}, \frac{1}{2} \right\} \left(1 - \frac{1}{2} - \frac{1}{4} \right), \tag{11}$$

which is valid for all $p \in [1, 2]$. In the case that $1 \leq p \leq 1 + 1/\log(2d)$, we note that $d^{-1/p^*} = 1/d^{\frac{p-1}{p}} \geq 1/e$, which yields

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \geq c \cdot \sqrt{\frac{\log(2d)}{n}} \wedge 1$$

for a numerical constant $c_0 > 0$.

To conclude, we need to establish the upper bound. Let us choose $a = 1 + 1/\log(2d)$, $\frac{\sup_{\theta \in \Theta} \|\theta\|_a \sup_{g \in \mathbf{B}_\gamma(0,1)} \|g\|_{a^*}}{\sqrt{a-1}\sqrt{n}}$ upper bounds the minimax regret. Since $a > p$, $\sup_{\theta \in \Theta} \|\theta\|_a = 1$. We have $a^* = \log(2d) + 1$ and $p^* \geq a^*$. We have

$$\|g\|_{a^*} \leq d^{\frac{1}{a^*} - \frac{1}{p^*}} \|g\|_{p^*} \leq d^{\frac{1}{a^*}},$$

because $g \in \mathbf{B}_{p^*}(0, 1)$. We note that $d^{1/a^*} = \exp\left(\frac{\log d}{\log(2d) + 1} \right) \leq e$. Noting that $1/\sqrt{2(a-1)} = \sqrt{\log(2d)/2}$ concludes this case. □

D.1.2 Case $1 + 1/\log(2d) < p \leq 2$

Let $d_0 \leq d$. We can embed a function $F_{d_0} : \mathbf{R}^{d_0} \times \mathcal{X} \to \mathbf{R}$ as a function $F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$ by letting π_{d_0} denote the projection onto the first d_0-components, and defining

$$F(\theta, x) = F_{d_0}(\pi_{d_0} \theta, x).$$

If the subgradients of F_{d_0} lie in $\mathbf{B}_{p^*}(0, 1)$, so do those of F. Similarly, if $\theta_0 \in \{\tau \in \mathbf{R}^{d_0}, \|\tau\|_p \leq 1\}$ then $\theta = (\theta_0, \mathbf{0}_{d_0+1:d}) \in \mathbf{B}_p(0, 1)$. As such, any lower bound for the d_0-dimensional problem implies an identical one for all $d \geq d_0$-dimensional problems. For $1 + 1/\log(2d) < p \leq 2$, let us define $d_0 = \lceil 1/2 \exp(\frac{1}{p-1}) \rceil$, so $d_0 \le d$ as desired. In the case that $p > 1 + 1/\log 16$, Lemma 3 yields the desired lower bound. In the case that $p \le 1 + 1/\log 16$, we have that $d_0 \ge 8$, and the lower bound (11) holds so that for a numerical constant $c > 0$,

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \ge c d_0^{-1/p^*} \cdot \sqrt{\frac{\log(2d_0)}{n}} \wedge 1.$$

We have that $d_0^{-1/p^*} \ge (1/2)^{\frac{1}{p}-1} \exp(-1/p) \ge \sqrt{2/e}$. This yields the final lower bound

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \ge c \cdot \frac{1}{\sqrt{2(p-1)n}} \wedge 1.$$

Proposition 5 yields the upper bound and concludes this proof. □

D.2 Proof of Theorem 4

Let $A \succ 0$ be a positive semi-definite matrix for the distance generating function $h_A(\theta) = \frac{1}{2}\theta^\top A \theta$ defined above, and let $q = \frac{p}{p-1}$ be the conjugate to p. We choose linear functions $F_i(\theta) := g_i^\top \theta$ where $g_i \in \mathbf{B}_q(0, 1)$. In this case, letting $\{\theta_i\}_{i \le n}$ be the points mirror descent plays, the regret with respect to $\theta \in \mathbf{R}^d$ is

$$\mathsf{Regret}_{n,A}(\theta) = \sum_{i \le n} F_i(\theta_i) - F_i(\theta) = \sum_{i \le n} g_i^\top (\theta_i - \theta),$$

so that

$$\mathsf{Regret}^*_{n,A} := \sup_{\|\theta\|_p \le 1} \mathsf{Regret}_{n,A}(\theta) = \left\| \sum_{i \le n} g_i \right\|_q + \frac{1}{2} \sum_{i \le n} \|g_i\|_{A^{-1}}^2 - \frac{1}{2} \left\| \sum_{i \le n} g_i \right\|_{A^{-1}}^2.$$

Now, we choose linear functions f_i so that the regret is large. To do so, choose vectors

$$u \in \operatorname*{argmax}_{\|x\|_q \le 1} x^\top A^{-1} x \ \text{ and } \ v \in \operatorname*{argmin}_{\|x\|_q = 1} x^\top A^{-1} x. \tag{12}$$

Now, we choose the vectors $g_i \in \mathbf{R}^d$ so that for a $\delta \in [0, 1]$ to be chosen,

(a) $g_i = u$ for $n/4$ of the indices $i \in [n]$

(b) $g_i = -u$ for $n/4$ of the indices $i \in [n]$

(c) $g_i = v$ for $\frac{n}{4}(1 + \delta)n$ of the indices $i \in [n]$

(d) $g_i = -v$ for $\frac{n}{4}(1 - \delta)$ of the indices $i \in [n]$.

With these choices, we obtain the regret lower bound

$$\begin{aligned} \mathsf{Regret}^*_{n,A} &\ge \sup_{\delta \le 1} \left[\frac{n}{2} \delta \|v\|_q + \frac{n}{4} u^\top A^{-1} u - \frac{\delta^2 n^2}{8} v^\top A^{-1} v \right] \\ &\ge \frac{n}{4} \cdot \left[u^\top A^{-1} u + \min\left\{1, \frac{2\|v\|_q}{n v^\top A^{-1} v}\right\} \|v\|_q \right]. \end{aligned} \tag{13}$$

We now consider two cases. In the first, A is large enough that $\|v\|_q \ge \frac{1}{2} n v^\top A^{-1} v$. Then the regret bound (13) becomes

$$\mathsf{Regret}^*_{n,A} \ge \frac{n}{4} \left[u^\top A^{-1} u + \|v\|_q \right] \ge \frac{n}{4},$$

as $\|v\|_q = 1$ by the construction (12). This gives the first result of the theorem. For the second claim, which holds in the case that $\|v\|_q < \frac{1}{2} n v^\top A^{-1} v$, we consider the operator norms of general invertible linear operators. For a mapping $T : \mathbf{R}^d \to \mathbf{R}^d$, define the ℓ_p to ℓ_q operator norm

$$\|T\|_{\ell_p \to \ell_q} := \sup_{x \ne 0} \frac{\|T(x)\|_q}{\|x\|_p}.$$

Then the construction (12) evidently yields

$$u^\top A^{-1}u = \|A^{-1/2}\|_{\ell_q\to\ell_2}^2 \quad \text{and} \quad \frac{\|v\|_q^2}{v^\top A^{-1}v} = \sup_{x\neq 0}\frac{\|A^{1/2}x\|_q^2}{\|x\|_2^2} = \|A^{1/2}\|_{\ell_2\to\ell_q}^2.$$

Revisiting the regret (13), we obtain

$$\mathsf{Regret}^*_{n,A} \geq \frac{n}{4}\cdot\left[\left\|A^{-1/2}\right\|^2_{\ell_q\to\ell_2} + \frac{2}{n}\left\|A^{1/2}\right\|^2_{\ell_2\to\ell_q}\right] \geq \sqrt{\frac{n}{2}}\|A^{-1/2}\|_{\ell_q\to\ell_2}\|A^{1/2}\|_{\ell_2\to\ell_q},$$

where we have used that $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for all a, b. But for any invertible linear operator, standard results on the Banach-Mazur distance [26, Corollary 2.3.2] imply that

$$\inf_{A\succ 0}\|A\|_{\ell_2\to\ell_q}\left\|A^{-1}\right\|_{\ell_q\to\ell_2} \geq d^{1/2-1/q}.$$

This gives the result. □

E Proof of Theorem 5

The proof follows similar lines as the one we show in Appendix D.2 but choosing different $u, v \in \mathbf{R}^d$. Let $\alpha \geq 0$ be a stepsize. We consider linear functions $F_i(\theta) := g_i^\top\theta$ with $\|\beta\odot g_i\|_1 \leq 1$. Let $\{\theta_i\}_{i\leq n}$ be the iterates of online gradient descent. The regret with respect to $\theta\in\mathbf{R}^d$ is

$$\mathsf{Regret}_{n,\alpha}(\theta) = \sum_{i\leq n} g_i^\top(\theta_i - \theta).$$

This yields

$$\mathsf{Regret}^*_{n,\alpha} = \sup_{\|\theta\|_\infty\leq 1}\mathsf{Regret}_{n,\alpha}(\theta) = \left\|\sum_{i\leq n} g_i\right\|_1 + \frac{\alpha}{2}\sum_{i\leq n}\|g_i\|_2^2 - \frac{\alpha}{2}\left\|\sum_{i\leq n} g_i\right\|_2^2.$$

Let $k = \arg\min_{j\leq d}\beta_j$, we choose

$$u = e_k/\beta_k \ \text{ and } \ v = \frac{\mathbf{1}}{\|\beta\|_1}.$$

For $\delta\in[0,1]$, we now choose the vectors $g_i\in\mathbf{R}^d$ as follows:

(a) $g_i = u$ for $n/4$ of the indices $i\in[n]$.
(b) $g_i = -u$ for $n/4$ of the indices $i\in[n]$.
(c) $g_i = v$ for $\frac{n}{4}(1+\delta)$ of the indices $i\in[n]$.
(d) $g_i = -v$ for $\frac{n}{4}(1-\delta)$ of the indices $i\in[n]$.

For this construction, we lower bound the regret

$$\begin{aligned}\mathsf{Regret}^*_{n,\alpha} &\geq \sup_{0\leq\delta\leq 1}\left\{\frac{n\delta}{2}\|v\|_1 + \frac{n\alpha}{4}\|u\|_2^2 - \frac{\alpha\delta^2 n^2}{8}\|v\|_2^2\right\} \\ &\geq \frac{n\alpha}{4}\|u\|_2^2 + \frac{n\|v\|_1}{4}\min\left\{1, \frac{2\|v\|_1}{n\alpha\|v\|_2^2}\right\}.\end{aligned} \tag{14}$$

If the stepsize is too small (i.e. $\alpha\leq\frac{2}{n}\frac{\|v\|_1}{\|v\|_2^2}$) then (14) becomes

$$\mathsf{Regret}^*_{n,\alpha} \geq \frac{nd}{4\|\beta\|_1}.$$

In the other case that $\alpha > \frac{2}{n}\frac{\|v\|_1}{\|v\|_2^2}$, (14) yields

$$\mathsf{Regret}^*_{n,\alpha} \geq \frac{n}{4\alpha}\|u\|_2^2 + \frac{\|v\|_1^2}{\|v\|_2^2}\frac{\alpha}{2} \geq \frac{\sqrt{2}}{2}\frac{\sqrt{nd}}{\min_{j\leq d}\beta_j},$$

which is the desired result. □