– Projects    – Recap of convergence, gen.
– SGD    – Tied down to a procedure.

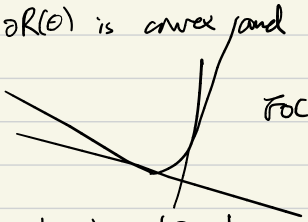
# Information Theoretic Lower Bounds

**Recap** We studied convex (stochastic) optimization problems

$$\underset{\theta \in \Theta}{\text{minimize}} \quad R(\theta), \quad \Theta \subseteq \mathbb{R}^d \text{ convex, compact}, \quad R: \Theta \to \mathbb{R} \text{ convex.}$$

<u>Def</u> (subgradients) The subgradient set of $R$ at $\theta$ is $\partial R(\theta) := \{g : R(\theta') \geq R(\theta) + \langle g, \theta' - \theta \rangle \ \forall \theta'\}$

cf. $\partial R(\theta)$ is convex and compact. subgradients are supporting hyperplanes of the epigraph



FOC: $\theta^* = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ iff $\langle g, \theta - \theta^* \rangle \geq 0 \quad \exists g \in \partial R(\theta^*) \ \forall \theta \in \Theta$

**Example** $h(\theta) = |\theta - c|$ $\qquad \partial h(\theta) = \begin{cases} \text{sign}(\theta - c) & \text{if } \theta \neq c \\ [-1, 1] & \text{if } \theta = c \end{cases}$

cf. Useful for e.g. $\min \frac{1}{n}\sum \ell(\theta; X, Y) + \lambda \|\theta\|_1$ or $\min \frac{1}{n}\sum |Y_i - \theta^T X_i|$

Stochastic (sub)gradient descent : Consider stochastic subgradients $g(\theta; \xi)$ st. for some independent RV $\xi$, $\mathbb{E}\, g(\theta; \xi) \in \partial R(\theta)$.

$$\theta^{k+1} \leftarrow \Pi_\Theta\left(\theta^k - \alpha_k g(\theta^k; \xi_k)\right) \quad : \text{SGD update}$$

e.g. $\xi_k = (X_k, Y_k)$. $\quad g(\theta; \xi) = \nabla_\theta \ell(\theta; X, Y)$. (or you can use batch $> 1$)

<u>Theorem</u> Let $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} R(\theta) > -\infty$, $D > 0$ s.t. $\sup_{\theta \in \Theta} \|\theta - \theta^*\|_2 \leq D$, $\exists M > 0$ s.t. $\sup_{\theta \in \Theta} \mathbb{E}\|g(\theta; \xi)\|_2^2 \leq M^2$.
Let $\alpha_k$ be dec, pos. step sizes, and $\bar{\theta}_K = \frac{1}{K}\sum_1^K \theta_k$

$$\mathbb{E}[R(\bar{\theta}_K) - R(\theta^*)] \leq \frac{D^2}{2K\alpha_K} + \frac{1}{2K}\sum_1^K \alpha_k M^2.$$

Pf) Identical as last week's result for differentiable functions.

<u>Cor</u> setting $\alpha_k = \frac{D}{M\sqrt{k}}$, $\quad \mathbb{E}\, R(\bar{\theta}_K) - R(\theta^*) \leq \frac{3DM}{2\sqrt{K}}$.

<u>Rmk</u> To solve problem to $\varepsilon$-accuracy, you need $O\left(\frac{1}{\varepsilon^2}\right)$-iterations of SGD.
Compared to $\log(1/\varepsilon)$ of IPMs, this is terribly slow.

<u>Q</u> Can we improve the SGD rate of conv? Or is it "optimal"? i.e. unimprovable

**Minimax rates** Considers worst-case optimality gap among a class of "problems".

<u>Components</u>
1) A collection of functions $\mathcal{F} := \{\theta \mapsto \mathbb{E}_P \ell(\theta; Z) : P \in \mathcal{P} \ \forall \ell\}$ induced by class of probabilities $\mathcal{P}$
2) closed, convex set $\Theta \subseteq \mathbb{R}^d$ — indep of everything

3) A stochastic gradient oracle $g : \mathbb{R}^d \times \Xi \times \mathcal{F} \to \mathbb{R}^d$ s.t. $\mathbb{E}\, g(\theta; \xi; R) \in \partial R(\theta)$.
Implicitly a prob. distribution on $\Xi$ induced by $P$.

↳ Think of this as # access to data points.

<u>Example</u>
$R_P(\theta) = \mathbb{E}_P \log(1 + \exp(-Y\theta^T X))$ : Logistic regression $\mathcal{P} = \{P \text{ s.t. } Y \in \{-1, 1\}, X \in [0,1]^d \text{ a.s.}\}$
$\mathcal{F} = \{R_P : P \in \mathcal{P}\}$ Define $\xi = (X, Y)$.
$g(\theta; \xi; P) = \nabla_\theta \ell(\theta; X, Y) = -\dfrac{YX}{1 + \exp(Y\theta^T X)}$, $\mathbb{E}_{\xi \sim P}\, g(\theta; \xi; P) = \nabla_\theta R_P(\theta)$.

View algorithms as making $K$ queries to the stoch. grad. oracle at $\theta_1, \cdots, \theta_K$ (adaptively) and it returns $\hat{\theta}_K$.

Optimality gap: $\mathbb{E}_{\xi_i}\left[R_P(\hat{\theta}(\xi_1, \cdots, \xi_K)) - \inf_{\theta \in \Theta} R_P(\theta)\right]$

↓ IDK what this is. So we think of nature choosing the worst-possible instance.

Measure performance w.r.t. hardest problem (dist. $P \in \mathcal{P}$)
ie, $\sup_{P \in \mathcal{P}} \mathbb{E}_{\xi_i}\left[R_P(\hat{\theta}(\xi_1, \cdots, \xi_K)) - \inf_{\theta \in \Theta} R_P(\theta)\right]$ ... (*)
i.e. I want $\hat{\theta}$ to achieve low error uniformly over $P \in \mathcal{P}$.

My job as statistical modeler / stoch optimizer is to come up with algo $\hat{\theta}$ s.t. (*) low.

The optimal algo gives the minimax risk
$$\mathfrak{M}_K(\Theta, \mathcal{P}) := \inf_{\hat{\theta}} \ \sup_{\forall \ell,\, P \in \mathcal{P}} \mathbb{E}_{\xi_i}\left[R_P(\hat{\theta}(\xi_1, \cdots, \xi_K)) - \inf_{\theta \in \Theta} R_P(\theta)\right]$$
where inf is over all measurable ftns of $\xi_1, \cdots, \xi_K$.

<u>Rmk</u> We can consider random procedures that incorporate more randomness using same techniques / bounds / proofs.

This measures extent to which we can optimize worst-case opt gap over the class of problems $P \in \mathcal{P}$, using finite access to stoch grad oracle.

<u>Rmk</u> Similar quantities can be defined for estimation problems where $\xi_1, \cdots, \xi_K$ are data points. This would measure sample complexity. Everything in today's class has a straightforward analogue to this setting.

<u>Rmk</u> Though we define minimax risk over $K$ queries to stoch grad oracle, we can see same bound applies to all procedures using $K$ samples.

**Rmk** Minimax can be very conservative.

**Step 0** (Worst-case → Bayesian) Let $\{P_v\} \subseteq \mathcal{P}$ be a collection of distributions indexed by finite or countable $\mathcal{V}$. Let $\pi$ be a prob on $\mathcal{V}$. For any fixed $\ell$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[R_P(\hat{\theta}_K) - \inf_{\theta \in \Theta} R_P(\theta)\right] \geq \sum_{v \in \mathcal{V}} \pi(v)\, \mathbb{E}\left[\underset{\overset{\shortparallel}{R_v}}{R_{P_v}}(\hat{\theta}_K) - \inf_{\theta \in \Theta} R_{P_v}(\theta)\right]$$

**Step 1** (Reduction from optimization to hypothesis testing)

**Def** For two cvx functions $R_0, R_1$ define the optimization separation bet $R_0$ & $R_1$ as

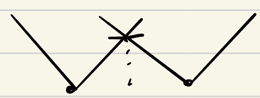
$$d_{opt}(R_0, R_1) := \sup\left\{\delta \geq 0 : \begin{array}{l} R_1(\theta) \leq R_1^* + \delta \Rightarrow R_0(\theta) \geq R_0^* + \delta \\ R_0(\theta) \leq R_0^* + \delta \Rightarrow R_1(\theta) \geq R_1^* + \delta \end{array} \text{ for any } \theta \in \Theta\right\}$$

where $R_v^* = \inf_{\theta \in \Theta} R_v(\theta)$.

↳ Optimizing $R_1$ means we can't optimize $R_0$ very well if $d_{opt}(R_0, R_1)$ large. This says if we optimized one function well, then we couldn't have optimized other functions well-separated in dopt.

e.g. If $\theta$ is s.t. $R_1(\theta) - R_1^* \leq d_{opt}(R_0, R_1)$, then $\theta$ cannot optimize $R_0$ well.

**Ex** $R_0(\theta) = |\theta - D|$ $\quad R_1(\theta) = |\theta + D|$



$d_{opt}(R_0, R_1) = D$.

Consider canonical hypothesis testing problem:

1) Nature chooses $V \in \mathcal{V}$ unif. at random

2) Cond. on $V = v$, we observe subgradients associated with $R_{P_v}$ for i.i.d $\xi_1, \ldots, \xi_K$.

**Goal** Figure out which index Nature chose.

If we can opt to better than $d_{opt}(R_v, R_{v'})$ $\forall v' \in \mathcal{V}$, then we can identify $V = v$.

**Lemma** Let $V$ be drawn u.a.r. from $\mathcal{V}$, $|\mathcal{V}| < \infty$, and assume $\{P_v\}_{v \in \mathcal{V}}$ is $\delta$-separated: $d_{opt}(R_v, R_{v'}) \geq \delta$ $\forall v \neq v' \in \mathcal{V}$. For any fixed $\ell$,

Then, $\displaystyle \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}\left[R_v(\hat{\theta}_K) - R_v^*\right] \geq \delta \inf_{\hat{v}} \underset{\text{over } V \,\&\, \xi_i^K}{\mathbb{P}}(\hat{v} \neq V)$

where inf is over all testing procedures based on observed data.

↳ In particular, this lower bounds Bayes-risk with $\pi(v) = \frac{1}{|\mathcal{V}|}$.

**Game plan** Construct a class of well-separated functions in dopt. Then show testing among them is difficult.

↳ Trade-off: Easier to distinguish functions when $\delta$ large, and vice-versa.

Pf) $\mathbb{E}[R_v(\hat\theta) - R_v^*] \geq \delta\, \mathbb{E}\,\mathbb{1}\{R_v(\hat\theta) - R_v^* \geq \delta\} = \delta\, \mathbb{P}_v(R_v(\hat\theta) - R_v^* \geq \delta)$. ✓ $\mathbb{P}_v$ is over $\xi^K$, for the oracle it defines.

Define the hypothesis test $\hat v = \begin{cases} v & \text{if } R_v(\hat\theta) \leq R_v^* + \delta \\ \text{random} & \text{o/w} \end{cases}$ ← this can be arbitrary

↳ this is well-defined since $d_{opt}(R_v, R_{v'}) \geq \delta$ $v \neq v'$, so such $v$ is unique.

Now note $\hat v \neq v \Rightarrow R_v(\hat\theta) \geq R_v^* + \delta$. So $\mathbb{P}_v(\hat v \neq v) \leq \mathbb{P}_v(R_v(\hat\theta) \geq R_v^* + \delta)$.

Hence, $\frac{1}{|V|}\sum_{v \in V} \mathbb{E}[R_v(\hat\theta_K) - R_v^*] \geq \delta \frac{1}{|V|}\sum_{v\in V} \mathbb{P}_v(\hat v \neq v) = \delta \cdot \mathbb{P}(\hat v \neq V)$ by def of $V$. ☒

**Le Cam's method.** Reduction to binary hypothesis testing. $\mathcal{V} = \{-1, +1\}$

We construct $P_1, P_{-1}$, and show hardness of optimizing on $\mathbb{R}^1 \supseteq \Theta$.

Def
TV distance $\|P - Q\|_{TV} = \sup_{A \subseteq \Xi \text{ meas.}} |P(A) - Q(A)|$
KL divergence $D_{KL}(P, Q) = \int_\Xi p(\xi) \log p(\xi)/q(\xi)\, d\mu(\xi)$

Lemma 1 $1 - \|P_1 - P_{-1}\|_{TV} = \inf_{\hat v} \{P_1(\hat v \neq 1) + P_{-1}(\hat v \neq -1)\}$ ← best prob of being wrong in binary hypo test,

Pf) Any $\hat v: \Xi \to \{-1, 1\}$, define $A = \hat v^{-1}\{1\}$, $A^c = \hat v^{-1}\{-1\}$ so $P_1(\hat v \neq 1) + P_{-1}(\hat v \neq -1) = P_1(A^c) + P_{-1}(A) = 1 - P_1(A) + P_{-1}(A)$
Taking inf over $\hat v$, RHS $= \inf_{A \subseteq \Xi \text{ meas}} \{1 - P_1(A) + P_{-1}(A)\} = 1 - \sup_{A \subseteq \Xi \text{ meas}} P_1(A) - P_{-1}(A) = 1 - \|P_1 - P_{-1}\|_{TV}$. ☒

We get $\mathfrak{M}_n(\Theta, \mathcal{P}) \geq \inf_{\hat\theta_K} \max_{v \in \{\pm 1\}} \mathbb{E}_{P_v}[R_v(\hat\theta_K) - R_v^*] \geq \delta \cdot \frac{1}{2} \inf_{\hat v} \{P_1^K(\hat v \neq 1) + P_{-1}^K(\hat v \neq -1)\}$

where $P_{\pm 1}^K$ are $K$-product distr over $\xi_1, \dots, \xi_K$. $= \delta \cdot \frac{1}{2} \cdot (1 - \|P_1^K - P_{-1}^K\|_{TV})$

↳ If $R_1, R_{-1}$ are diff, then $\delta$ is big (optimizing one makes other worse), but this will likely make $P_1, P_{-1}$ to be different.

Now, $\|P_1^K - P_{-1}^K\|_{TV}$ is unwieldy since TV distance don't play well with products.

Lemma 2 (Pinsker) $\|P - Q\|_{TV}^2 \leq \frac{1}{2} D_{KL}(P, Q)$

Lemma 3 $D_{KL}(P_1^K, P_{-1}^K) = K \cdot D_{KL}(P_1, P_{-1})$

Pf) $\int \prod_{k=1}^K p_1(\xi_k) \log \frac{\prod_{k=1}^K p_1(\xi_k)}{\prod_{k=1}^K p_{-1}(\xi_k)} d\xi_1 \cdots d\xi_K = \sum_{k=1}^K \int \left(\prod_{k=1}^K p_1(\xi_k)\right) \log \frac{p_1(\xi_k)}{p_{-1}(\xi_k)} d\xi_1 \cdots d\xi_K$

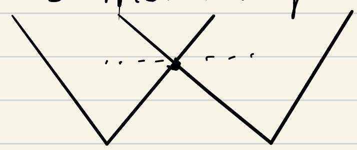
$= \sum_{k=1}^K \int p_1(\xi_k) \log \frac{p_1(\xi_k)}{p_{-1}(\xi_k)} d\xi_k = \sum_{k=1}^K D_{KL}(P_1, P_{-1})$ ☒.

So we get $\|P_1^K - P_{-1}^K\|_{TV} \leq \sqrt{\frac{K}{2} D_{KL}(P_1, P_{-1})}$. Plugging this in,

$$\mathfrak{M}_K(\Theta, \mathcal{P}) \geq \frac{1}{2} d_{opt}(R_1, R_{-1}) \cdot \left(1 - \sqrt{\frac{K}{2} D_{KL}(P_1, P_{-1})}\right).$$ ☒

Now, construct $P_1, P_{-1}$ (with corresponding oracles) s.t. they're close in KL so testing is hard, but well-separated in $d_{opt}$.

Let $\Theta \subseteq \mathbb{R}^d$ be s.t. $L^2$-ball of radius $D$ is included in $\Theta$. We consider stoch. grad. oracles s.t. $\mathbb{E}\|g(\theta;\xi;R_P)\|_2^2 \leq M^2$ for all $\theta \in \Theta$. Let $\mathcal{P}$ be s.t. $\theta \mapsto R_P(\theta)$ is $M$-Lip in $\|\cdot\|_2$.



**Construction** Consider $d=1$ w.l.o.g.: Assume $[-D,D] \subseteq \Theta \subseteq \mathbb{R}$.

Define $P_1, P_{-1}$ s.t. $R_1(\theta) = \delta M|\theta - D|$ $\quad R_{-1}(\theta) = \delta M|\theta + D|$.

From picture, $d_{opt}(R_1, R_{-1}) = \delta M D$.

Now, construct stoch. grad. oracle s.t. $\mathbb{E}\, g(\theta;\xi;R_v)^2 \leq M^2$.

For $\delta < 1$, the oracle for $P_v$, $v \in \{\pm 1\}$ is

flip coin $\xi \sim \mathrm{Ber}((1+v\delta)/2)$, $g(\theta;\xi;R_v) = \begin{cases} vM\,\mathrm{sgn}(\theta - vD) & \text{if } \xi = 1 \\ -vM\,\mathrm{sgn}(\theta - vD) & \text{if } \xi = 0 \end{cases}$

cf. If $\theta = vD$, return random $\pm M$.

$$\mathbb{E}_{P_v} g(\theta;\xi;R_v) = \frac{1+v\delta}{2} vM\,\mathrm{sgn}(\theta - vD) - \frac{1-v\delta}{2} vM\,\mathrm{sgn}(\theta - vD) = v\delta \cdot vM\,\mathrm{sgn}(\theta - vD) = \delta M\,\mathrm{sgn}(\theta - vD) \in \partial R_v(\theta)$$

**Distance**
$$D_{KL}(P_1, P_{-1}) = \sum_{\xi \in \{0,1\}} P_1(\xi) \log \frac{P_1(\xi)}{P_{-1}(\xi)} = \frac{1+\delta}{2} \log \frac{\frac{1+\delta}{2}}{\frac{1-\delta}{2}} + \frac{1-\delta}{2} \log \frac{\frac{1-\delta}{2}}{\frac{1+\delta}{2}}$$
$$= \delta \log \frac{1+\delta}{1-\delta}$$

Taylor expansion of $x \mapsto \log \frac{1+x}{1-x} = a(x)$ at $x=0$: $\quad a'(x) = \frac{1}{1+x} + \frac{1}{1-x} \quad a''(x) = -\frac{1}{(1+x)^2} + \frac{1}{(1-x)^2}$

$$\log \frac{1+\delta}{1-\delta} = a(0) + a'(0)\delta + \frac{a''(\tilde{\delta})\delta^2}{2} \quad \text{for some } \tilde{\delta} \in [0,\delta]$$
$$\leq 0 + 2\delta + 2\delta^2 \quad \text{if } \delta \leq \tfrac{1}{2} \qquad \left(\because a''(\tilde{\delta}) \leq \frac{1}{(1-\tilde{\delta})^2} \leq \frac{1}{(1-\frac{1}{2})^2} = 4\right)$$
$$\leq 3\delta \quad \text{if } \delta \leq \tfrac{1}{2}.$$

So $D_{KL}(P_1, P_{-1}) \leq 3\delta^2$. Plugging this into Le Cam's result,

$$\mathcal{M}_K(\Theta, \mathcal{P}) \geq \frac{1}{2} \delta M D \cdot \left(1 - \sqrt{\frac{K}{2} 3\delta^2}\right).$$

**Set $\delta$** I need $1 - \sqrt{\frac{K}{2}\delta^2 \cdot 3}$ to behave like a const. i.e. $\delta^2 \sim \frac{1}{K}$.

Set $\delta = \frac{1}{\sqrt{6K}}$ so that $1 - \sqrt{\frac{3K}{2}\delta^2} = \frac{1}{2}$ $\quad (K \geq 2)$.

Conclude $\mathcal{M}_K(\Theta, \mathcal{P}) \geq \frac{MD}{4\sqrt{6K}}$. for $K \geq 2$.

Recalling prev result, if $D_{inner}\mathbb{B}_2 \subset \Theta \subseteq D_{outer}\mathbb{B}_2$,

$$\frac{D_{inner} M}{\sqrt{K}} \lesssim \mathcal{M}_K(\Theta, \mathcal{P}) \lesssim \frac{D_{outer} M}{\sqrt{K}}.$$

**Q** My construction was in $\mathbb{R}^1$. Surely things are harder in $d > 1$. Can we show a more explicit bound with dimension?

# Assouad's method

Reduce opt to $d$ binary hypothesis tests.

Let $\mathcal{V} = \{\pm 1\}^d$ be the $d$-dim binary hypercube. For each $v \in \mathcal{V}$, we construct function $R_v$ and $P_v$, a joint distribution over $\xi_1^k$.

W.l.o.g. assume $0 \in \Theta$. We say $\{P_v\}_{v\in\mathcal{V}}$ is $\vec{\delta}$-separated in the Hamming distance if for all $\theta \in \Theta$
$R_v(\theta) - R_v^* \geq \sum_{j=1}^d \delta_j \mathbb{1}\{\mathrm{sgn}(\theta_j) \neq v_j\}$.

Ex $R_v(\theta) = \sum_{j=1}^d \delta_j |\theta_j - v_j|$, $\mathcal{V} \subseteq \Theta$ so $R_v^* = 0 \ \forall v \in \mathcal{V}$. Then, $R_v(\theta) - R_v^* = \sum_{j=1}^d \delta_j |\theta_j - v_j| \geq \sum_{j=1}^d \delta_j \mathbb{1}\{\mathrm{sgn}(\theta_j) \neq v_j\}$

**Lemma** (Assouad) Let $\{P_v\}_{v\in\mathcal{V}}$ be $\vec{\delta}$-separated in Hamming distance, $\mathcal{V} = \{\pm 1\}^d$. Let $P_{\pm j} = \frac{1}{2^{d-1}} \sum_{v_j = \pm 1} P_v$

$$\frac{1}{2^n} \sum_{v \in \{\pm 1\}^d} \mathbb{E}\, R_v(\hat{\theta}_k) - R_v^* \geq \frac{1}{2} \sum_1^d \delta_j \left(1 - \|P_{+j} - P_{-j}\|_{TV}\right)$$

Pf)
$$\frac{1}{2^d} \sum_{v\in\{\pm1\}^d} \mathbb{E}\left[R_v(\hat{\theta}_k) - R_v^*\right] \geq \frac{1}{2^d} \sum_{v\in\{\pm1\}^d} \sum_{j=1}^d \delta_j\, \mathbb{P}_v\left(\mathrm{sgn}(\hat{\theta}_j) \neq v_j\right) = \sum_{j=1}^d \delta_j \frac{1}{2^d} \sum_{v\in\{\pm1\}^d} \mathbb{P}_v\left(\mathrm{sgn}(\hat{\theta}_j) \neq v_j\right)$$
$$= \sum_1^d \delta_j \cdot \frac{1}{2^d} \left\{ \sum_{v_j = 1} P_v\left(\mathrm{sgn}(\hat{\theta}_j) \neq 1\right) + \sum_{v_j = -1} P_v\left(\mathrm{sgn}(\hat{\theta}_j) \neq -1\right) \right\}$$
$$= \frac{1}{2} \sum_1^d \delta_j \left\{ P_{+j}\left(\mathrm{sgn}(\hat{\theta}_j) \neq 1\right) + P_{-j}\left(\mathrm{sgn}(\hat{\theta}_j) \neq -1\right) \right\}$$
$$\geq \frac{1}{2} \sum_1^d \delta_j \left(1 - \|P_{+j} - P_{-j}\|_{TV}\right) \quad \boxtimes.$$

**Cor** Let $P_{v,\pm j}$ be $P_{v'}$ where $v'$ is $v$ with $j^{\text{th}}$ entry of $v$ forced to be $\pm 1$. Then,
$$\mathfrak{M}_K(\Theta, \mathcal{P}) \geq \frac{d\delta}{2}\left(1 - \sqrt{\max_{v\in\mathcal{V},\, 1\leq j\leq d} D_{KL}(P_{v,+j}, P_{v,-j})/2}\right)$$
whenever $\delta \cdot \mathbb{1}$-separation in Hamming holds.

Pf) Note that $P_{\pm j} = \frac{1}{2^d} \sum_v P_{v,\pm j}$ since $\sum_v P_{v,+j} = \sum_{v: v_j = 1} P_v + \sum_{v: v_j = -1} P_{v,+j} = 2 \sum_{v: v_j = 1} P_v$

Then from triangle inequality,

$$\|P_{+j} - P_{-j}\|_{TV} = \left\|\frac{1}{2^d} \sum_v (P_{v,+j} - P_{v,-j})\right\|_{TV} \leq \frac{1}{2^d} \sum_v \|P_{v,+j} - P_{v,-j}\|_{TV} \leq \max_{v,j} \|P_{v,j} - P_{v,-j}\|_{TV} \leq \max_{v,j} \sqrt{\tfrac{1}{2} D_{KL}(P_{v,j}, P_{v,-j})}$$
by Pinsker.

$$\mathfrak{M}_K(\Theta, \mathcal{P}) \geq \frac{d\delta}{2}\left(1 - \frac{1}{d}\sum_1^d \|P_{+j} - P_{-j}\|_{TV}\right) \geq \frac{d\delta}{2}\left(1 - \max_{v,j} \sqrt{\tfrac{1}{2} D_{KL}(P_{v,j}, P_{v,-j})}\right). \quad \boxtimes.$$

Consider $[-D, D]^d \subseteq \Theta$, and $\mathbb{E}\|g(\theta; \xi; R)\|_1^2 \leq M^2 \ \forall \theta \in \Theta$, and $|R_P(\theta) - R_P(\theta')| \leq M \|\theta - \theta'\|_\infty$

Construction

Let $R_v(\theta) = \frac{M\delta}{d} \|\theta - D\cdot v\|_1$ for $v \in \{\pm 1\}^d$. Then

$R_v(\theta) - R_v^* = R_v(\theta) \geq \frac{M\delta}{d} \sum_1^d D \mathbb{1}\{\mathrm{sgn}(\theta_j) \neq v_j\}$. So $\frac{DM\delta}{d}\mathbb{1}$-separated in the Hamming distance.

For stoch. grad. oracle, take $\xi_v = \begin{cases} e_j & \text{w.p. } \frac{1+v_j\delta}{2d} \\ -e_j & \text{w.p. } \frac{1-v_j\delta}{2d} \end{cases} \quad \forall j = 1, \ldots, d$

$$g(\theta; \xi; R_v) = \begin{cases} M\cdot v_j \cdot \mathrm{sgn}(\theta_j - Dv_j)\cdot e_j & \text{if } \xi_v = e_j \\ -M v_j\, \mathrm{sgn}(\theta_j - Dv_j)\, e_j & \text{if } \xi_v = -e_j \end{cases}, \quad \text{I.M randomly if } \theta_j = Dv_j$$

$$\mathbb{E}_{P_v} g(\theta; \xi; R_v) = \sum_{j=1}^d \left\{ \frac{1+v_j\delta}{2d} M v_j\, \mathrm{sgn}(\theta_j - Dv_j) e_j - \frac{1-v_j\delta}{2d} M v_j\, \mathrm{sgn}(\theta_j - Dv_j) e_j \right\}$$
$$= \sum_{j=1}^d \frac{\delta M}{d} v_j^2\, \mathrm{sgn}(\theta_j - Dv_j) e_j = \frac{\delta M}{d}\, \mathrm{sgn}(\theta - Dv) \in \partial R_v(\theta)$$

**Distance** We need to bound $D_{KL}(P_v, P_{v'})$ for $v$ & $v'$ that differ only in a single coordinate.
W.l.o.g. let this be the first coordinate, and recall $P_v$ is a joint distribution over $\xi_1^K$.
Since $P_v$ is a product distribution over i.i.d. $\xi_i$, let $\xi_1 \sim P_{v,\xi_1}$. Then $P_v = P_{v,\xi_1}^K$.
So $D_{KL}(P_v, P_{v'}) = D_{KL}(P_{v,\xi_1}^K, P_{v',\xi_1}^K) = K\, D_{KL}(P_{v,\xi_1}, P_{v',\xi_1})$

Now, note from definition that if $v_j = v_j'$ $\forall j \neq 1$, with $v_1 = 1$, $v_1' = -1$

$$D_{KL}(P_{v,\xi_1}, P_{v',\xi_1}) = P(\xi_v = e_1) \cdot \log \frac{P(\xi_v = e_1)}{P(\xi_{v'} = e_1)} + P(\xi_v = -e_1) \log \frac{P(\xi_v = -e_1)}{P(\xi_{v'} = -e_1)}$$

$$= \frac{1+\delta}{2d} \log \frac{\frac{1+\delta}{2d}}{\frac{1-\delta}{2d}} + \frac{1-\delta}{2d} \log \frac{\frac{1-\delta}{2d}}{\frac{1+\delta}{2d}} = \frac{\delta}{d} \log \frac{1+\delta}{1-\delta} \le \frac{3}{d}\delta^2, \quad \delta \le \frac{1}{2}$$

So we conclude $\mathcal{M}_n(\Theta, \mathcal{P}) \ge \frac{d}{2} \cdot \frac{MD\delta}{d} \cdot \left(1 - \sqrt{\frac{1}{2} \cdot \frac{3K}{d} \delta^2}\right) = \frac{MD\delta}{2} \cdot \left(1 - \sqrt{\frac{3}{2d} K\delta^2}\right)$.
Setting $\delta^2 = \frac{d}{6K}$, which is possible if $\frac{d}{6} \le \frac{1}{2} \cdot K$, $\sqrt{\frac{3}{2d} K\delta^2} = \frac{1}{2}$.

$$\mathcal{M}_K(\Theta, \mathcal{P}) \ge \frac{MD}{2} \cdot \sqrt{\frac{d}{6K}} \cdot \frac{1}{2} = \frac{MD}{4}\sqrt{\frac{d}{6K}} \qquad \text{whenever} \quad K \ge \frac{d}{3}.$$

Let's say $[-D_{in}, D_{in}]^d \subseteq \Theta \subseteq [D_{out}, D_{out}]^d$. Then,

$$\frac{D_{in} \cdot M}{4}\sqrt{\frac{d}{6K}} \le \mathcal{M}_K(\Theta, \mathcal{P}) \le \frac{3}{2} D_{out} \cdot M \cdot \sqrt{\frac{d}{K}} \qquad \text{from previous result.}$$