Stats 231 / CS229T Homework 2 Solutions

Question 1 (Rademacher and Gaussian complexity)**:** In some situations it may be easier to control the *Gaussian complexity* of a set of functions than the Rademacher complexity. Given points $x_1, \ldots, x_n$, the (unnormalized) empirical Gaussian complexity is

$$\widehat{G}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n g_i f(x_i) \mid x_{1:n}\right]$$

where $g_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$ are independent standard Gaussians. The Gaussian complexity is the expected version of the empirical complexity $G_n(\mathcal{F}) = \mathbb{E}[\widehat{G}_n(\mathcal{F})]$. Show that, assuming that $\mathcal{F}$ is symmetric in the sense that if $f \in \mathcal{F}$ then $-f \in \mathcal{F}$,

$$n\widehat{R}_n(\mathcal{F}) \le \sqrt{\frac{\pi}{2}}\widehat{G}_n(\mathcal{F}).$$

Answer: Let ϵ_i denote a Rademacher random variable, taking values uniformly in $\{-1, +1\}$. We use the fact that $\epsilon_i |g_i| \sim \mathsf{N}(0,1)$, where $g_i \sim \mathsf{N}(0,1)$, and that

$$\mathbb{E}[|g_i|] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |g| \exp\left\{-g^2/2\right\} dg = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} g \exp\left\{-g^2/2\right\} dg = \sqrt{2/\pi}. \tag{1}$$

Then

$$\begin{aligned}
\widehat{G}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n g_i f(x_i) \middle| x_{1:n}\right] \\
&= \mathbb{E}_\epsilon\left[\mathbb{E}_g\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i |g_i| f(x_i) \middle| \epsilon_{1:n}, x_{1:n}\right]\right] \\
&\stackrel{(i)}{\ge} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \mathbb{E}[|g_i|] \epsilon_i f(x_i)\right] \\
&\stackrel{(ii)}{=} \sqrt{\frac{2}{\pi}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i)\right] \\
&\stackrel{(iii)}{=} \sqrt{\frac{2}{\pi}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] \\
&= \sqrt{\frac{2}{\pi}} n\widehat{R}_n(\mathcal{F}),
\end{aligned}$$

where (i) is from Jensen's inequality applied to the inner expectation and the convex supremum function, (ii) is from (1), and (iii) is by the symmetry of $\mathcal{F}$. □

Question 2 (Gaussian comparisons and contractions)**:** The *Sudakov-Fernique* bound is a comparison inequality for Gaussian processes that allows substantial control over Gaussian processes, including more powerful contraction inequalities than are available for Rademacher complexities. Recall that a collection $\{X_t\}_{t \in T}$ of random variables is a *Gaussian process* if X_t is normally distributed for all T and all pairs (X_t, X_s), where $s, t \in T$, are jointly normally distributed. Let

$\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ be Gaussian processes indexed by a set T.[1] The Sudakov-Fernique inequality is that if

$$\mathbb{E}[X_t] = \mathbb{E}[Y_t] = 0 \;\text{ and }\; \mathbb{E}[(X_t - X_s)^2] \le \mathbb{E}[(Y_t - Y_s)^2] \;\text{ for all } s, t \in T \tag{2}$$

then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

This is perhaps intuitive: the condition (2) suggests that X_t is somehow more tightly correlated with itself than Y_t, so that we expect Y_t to be "bigger" in some way.

(a) Prove *Slepian's inequality* from the Sudakov-Fernique bound. Slepian's inequality is that

$$\mathbb{E}[X_t X_s] \ge \mathbb{E}[Y_t Y_s] \;\text{ and }\; \mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2] \;\text{ for all } s, t \in T$$

implies $\mathbb{E}[\sup_{t\in T} X_t] \le \mathbb{E}[\sup_{t\in T} Y_t]$.

Now, let us use the Sudakov-Fernique condition (2) to give contraction inequalities for Gaussian complexity.

(b) Let $\phi_i : \mathbb{R}^d \to \mathbb{R}$ be M_i-Lipschitz for $i = 1, 2, \ldots, n$. Let $g_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ be independent standard Gaussians and $Z_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ be independent $\mathbb{R}^d$-valued Gaussian vectors with identity covariance. Define the empirical Gaussian complexities

$$\widehat{G}_n(\phi \circ \Theta) := \mathbb{E}\left[\sup_{\theta\in\Theta} \sum_{i=1}^n g_i \phi_i(\theta)\right] \;\text{ and }\; \widehat{G}_n(\Theta) := \mathbb{E}\left[\sup_{\theta\in\Theta} \sum_{i=1}^n M_i Z_i^T \theta\right].$$

Show that for a numerical constant $C < \infty$ (specify your constant)

$$\widehat{G}_n(\phi \circ \Theta) \le C \cdot \widehat{G}_n(\Theta).$$

(c) Let $\ell : \Theta \times \mathbb{R}^d \to \mathbb{R}$ satisfy $\ell(\theta, x) = \phi(\theta^T x)$ where ϕ is M-Lipschitz. Define $\mathcal{F}$ to be the loss class $\mathcal{F} := \{\ell(\theta, \cdot) : \theta \in \Theta\}$. Show that

$$\widehat{G}_n(\mathcal{F}) \le \widehat{G}_n(\Theta) := M\mathbb{E}\left[\sup_{\theta\in\Theta} \sum_{i=1}^n g_i \theta^T x_i\right]$$

(d) Fix $\theta^\star \in \Theta \subset \mathbb{R}^d$, and suppose that we instead use the centered loss class

$$\mathcal{F} := \{\ell(\theta, \cdot) - \ell(\theta^\star, \cdot) \mid \theta \in \Theta\}.$$

In addition, let $\Theta_\epsilon = \{\theta \in \Theta \mid \|\theta - \theta^\star\|_2 \le \epsilon\}$. Under the conditions of part (c), give an explicit upper bound on

$$\widehat{G}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{\theta\in\Theta_\epsilon} \sum_{i=1}^n g_i(\ell(\theta; x_i) - \ell(\theta^\star; x_i))\right].$$

What is your bound's dependence on ϵ, the Lipschitz constant M, n, and the dimension d of Θ? How does this compare to the localized Rademacher complexity result we gave in class?

Answer:

[1] Technically T must be finite, but in our settings we can approximate T by finite subsets so that everything holds.

(a) Assume $\mathbb{E}[X_t] = \mathbb{E}[Y_t] = 0$ for all t, which is all that is needed for this problem. It is easy to see that the Slepian assumption implies the Sudakov-Fernique assumption:

$$\mathbb{E}[(X_t - X_s)^2] = \mathbb{E}[X_t^2] - 2\mathbb{E}[X_t X_s] + \mathbb{E}[X_s^2] \leq \mathbb{E}[Y_t^2] - 2\mathbb{E}[Y_t Y_s] + \mathbb{E}[Y_s^2] = \mathbb{E}[(Y_t - Y_s)^2].$$

Therefore we can apply the Sudakov-Fernique inequality and conclude $\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[\sup_{t\in T} Y_t]$.

(b) Let X_θ be the mean-zero Gaussian process defined by $X_\theta = \sum_{i=1}^n g_i \phi_i(\theta)$, and similarly define $Y_\theta = \sum_{i=1}^n M_i Z_i^T \theta$. We must verify the Sudakov-Fernique condition for X_θ and Y_θ. We compute

$$\begin{aligned}
\mathbb{E}[(X_{\theta_1} - X_{\theta_2})^2] &= \mathbb{E}\left[\left(\sum_{i=1}^n g_i(\phi_i(\theta_1) - \phi_i(\theta_2))\right)^2\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\sum_{i=1}^n g_i^2(\phi_i(\theta_1) - \phi_i(\theta_2))^2\right] \\
&\overset{(ii)}{\leq} \sum_{i=1}^n \mathbb{E}[g_i^2] M_i^2 \|\theta_1 - \theta_2\|^2 \\
&= \|\theta_1 - \theta_2\|^2 \sum_{i=1}^n M_i^2,
\end{aligned}$$

where (i) is because g_i are uncorrelated so the cross terms disappear and (ii) is from the M_i-Lipschitzness of ϕ_i. Similarly,

$$\mathbb{E}(Y_{\theta_1} - Y_{\theta_2})^2] = \mathbb{E}\left[\left(\sum_{i=1}^n M_i Z_i^T(\theta_1 - \theta_2)\right)^2\right] = \|\theta_1 - \theta_2\|^2 \sum_{i=1}^n M_i^2.$$

So the condition is satisfied, and the conclusion follows from the Sudakov-Fernique inequality. The constant is $C = 1$.

(c) In a similar manner to part (b), define $X_\theta = \sum_{i=1}^n g_i \phi(\theta^T x_i)$ and $Y_\theta = M \sum_{i=1}^n g_i \theta^T x_i$. Then by the Lipschitz assumption,

$$\begin{aligned}
\mathbb{E}[(X_{\theta_1} - X_{\theta_2})^2] &= \mathbb{E}\left[\left(\sum_{i=1}^n g_i(\phi_i(\theta_1^T x_i) - \phi_i(\theta_2^T x_i))\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n g_i^2(\phi_i(\theta_1^T x_i) - \phi_i(\theta_2^T x_i))^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n g_i^2 M^2(\theta_1 - \theta_2)^T x_i)^2\right] \\
&= M^2 \sum_{i=1}^n ((\theta_1 - \theta_2)^T x_i)^2
\end{aligned}$$

We also compute

$$\mathbb{E}[(Y_{\theta_1} - Y_{\theta_2})^2] = \mathbb{E}\left[\left(M \sum_{i=1}^n g_i(\theta_1 - \theta_2)^T x_i\right)^2\right] = M^2 \sum_{i=1}^n ((\theta_1 - \theta_2)^T x_i)^2.$$

So the condition $\mathbb{E}[(X_{\theta_1} - X_{\theta_2})^2] \leq \mathbb{E}[(Y_{\theta_1} - Y_{\theta_2})^2]$ is satisfied, and the conclusion follows from the Sudakov-Fernique inequality.

(d) Applying the result from part (c), we find

$$\widehat{G}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{\theta \in \Theta_\epsilon} \sum_{i=1}^n g_i(\ell(\theta; x_i) - \ell(\theta^\star; x_i))\right] \leq M\mathbb{E}\left[\sup_{\theta \in \Theta_\epsilon} \sum_{i=1}^n g_i(\theta - \theta^\star)^T x_i\right]$$
$$= M\mathbb{E}\left[\sup_{\theta \in \Theta_\epsilon} \epsilon \left\|\sum_{i=1}^n g_i x_i\right\|_2\right] \stackrel{(i)}{\leq} M\epsilon\sqrt{\mathbb{E}\left[\sum_{i=1}^n \|g_i x_i\|_2^2\right]} = M\epsilon\sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2}$$

where inequality (i) is Jensen's inequality and uses the fact that $\mathbb{E}[g_i g_j x_i^T x_j] = 0$ for $i \neq j$.

In the case that the data x_i are bounded in ℓ_2-norm, by (say) r this yields

$$\widehat{G}_n(\mathcal{F}) \leq Mr\sqrt{n}\epsilon,$$

which is tighter than the local Rademacher complexity results we have given (which grow with dimension, as $\widehat{R}_n(\mathcal{F}) \leq M\epsilon\sqrt{\frac{d}{n}}$).

□

Question 3 (Adaptive stepsizes)**:** Consider an online learning problem in which we receive a sequence of convex functions $f_t : X \to \mathbb{R}$, where $X \subset \mathbb{R}^d$ is a compact convex set. Let $D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle$ be the usual Bregman divergence, and assume that

$$D_h(x, y) \leq D_X^2 \quad \text{for all } x, y \in X.$$

As usual, we define the regret of a sequence of plays $x_1, x_2, \ldots$ by

$$\mathsf{Reg}_T := \sum_{t=1}^T [f_t(x_t) - f_t(x^\star)]$$

where $x^\star \in \operatorname{argmin}_{x \in X} \sum_{t=1}^T f_t(x)$. We consider the usual online mirror descent algorithm

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{\langle g_t, x\rangle + \frac{1}{\alpha_t} D_h(x, x_t)\right\} \quad \text{where} \quad g_t \in \partial f_t(x_t).$$

Assume that $h : X \to \mathbb{R}$ is strongly convex with respect to the norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, so that $D_h(x, y) \geq \frac{1}{2}\|x - y\|^2$ for all $x, y \in X$.

(a) Show that for *any* (nonnegative) sequence of non-increasing stepsizes $\alpha_1, \alpha_2, \ldots$, we have

$$\mathsf{Reg}_T = \sum_{t=1}^T [f_t(x_t) - f_t(x^\star)] \leq \frac{D_X^2}{\alpha_T} + \sum_{t=1}^T \frac{\alpha_t}{2}\|g_t\|_*^2.$$

(b) Suppose that we choose a fixed stepsize $\alpha_t \equiv \alpha$ for all t. Give the value of

$$\inf_{\alpha \geq 0}\left\{\sum_{t=1}^T \frac{D_X^2}{\alpha} + \sum_{t=1}^T \frac{\alpha}{2}\|g_t\|_*^2\right\}.$$

(c) Let $\{a_t\}_{t=1}^T$ be an arbitrary sequence of non-negative numbers. Define $b_t = \sum_{\tau=1}^t a_\tau$. Prove that

$$\sum_{t=1}^T \frac{a_t}{\sqrt{b_t}} \leq 2\sqrt{b_T} = 2\sqrt{\sum_{t=1}^T a_t},$$

where we treat $0/0$ as 0.

(d) Based on parts (b) and (c), give a sequence of stepsizes α_t, which depend only on the subgradients $\{g_\tau\}_{\tau=1}^t$ through time t and the diameter D_X, such that

$$\frac{D_X^2}{\alpha_T} + \sum_{t=1}^T \frac{\alpha_t}{2} \|g_t\|_*^2 \leq O(1) \cdot \inf_{\alpha \geq 0} \left\{ \frac{D_X^2}{\alpha} + \frac{\alpha}{2} \sum_{t=1}^T \|g_t\|_*^2 \right\}.$$

Answer:

(a) This is the standard regret bound. We have

$$\begin{aligned}
\mathsf{Reg}_T &= \sum_{t=1}^T [f_t(x_t) - f_t(x^\star)] \overset{(i)}{\leq} \sum_{t=1}^T \langle g_t, x_t - x^\star \rangle \\
&= \sum_{t=1}^T \langle g_t, x_{t+1} - x^\star \rangle + \sum_{t=1}^T \langle g_t, x_t - x_{t+1} \rangle \\
&\overset{(ii)}{\leq} \sum_{t=1}^T \frac{1}{\alpha_t} \langle \nabla h(x_{t+1}) - \nabla h(x_t), x^\star - x_{t+1} \rangle + \sum_{t=1}^T \langle g_t, x_t - x_{t+1} \rangle
\end{aligned}$$

where the inequality (i) used convexity and inequality (ii) used that

$$\left\langle g_t + \frac{1}{\alpha_t}[\nabla h(x_{t+1}) - \nabla h(x_t)], y - x_{t+1} \right\rangle \geq 0 \quad \text{for all } y \in X$$

by the standard optimality conditions for convex problems and definition of the update for x_{t+1}.

Now we use our standard Bregman divergence identity that

$$\langle \nabla h(z) - \nabla h(x), y - z \rangle = D_h(y, x) - D_h(y, z) - D_h(z, x)$$

applied with $y = x^\star$, $z = x_{t+1}$, and $x = x_t$ to obtain the following upper bound on the regret:

$$\mathsf{Reg}_T \leq \sum_{t=1}^T \frac{1}{\alpha_t} \left[D_h(x^\star, x_t) - D_h(x^\star, x_{t+1}) - D_h(x_{t+1}, x_t) \right] + \sum_{t=1}^T \langle g_t, x_t - x_{t+1} \rangle .$$

Using the Fenchel-Young inequality as in class, we have

$$\langle g_t, x_t - x_{t+1} \rangle \leq \frac{\alpha_t}{2} \|g_t\|_*^2 + \frac{1}{2\alpha_t} \|x_t - x_{t+1}\|^2 \leq \frac{\alpha_t}{2} \|g_t\|_*^2 + \frac{1}{\alpha_t} D_h(x_{t+1}, x_t),$$

which gives us the bound

$$\begin{aligned}
\mathsf{Reg}_T &\le \sum_{t=1}^T \frac{1}{\alpha_t}\left[D_h(x^\star, x_t) - D_h(x^\star, x_{t+1})\right] + \sum_{t=1}^T \frac{\alpha_t}{2}\left\|g_t\right\|_*^2 \\
&= \sum_{t=2}^T \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right) D_h(x^\star, x_t) + \frac{1}{\alpha_1} D_h(x^\star, x_1) - \frac{1}{\alpha_T} D_h(x^\star, x_T) + \sum_{t=1}^T \frac{\alpha_t}{2}\left\|g_t\right\|_*^2 \\
&\stackrel{(iii)}{\le} \sum_{t=2}^T \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right) D_X^2 + \frac{1}{\alpha_1} D_X^2 + \sum_{t=1}^T \frac{\alpha_t}{2}\left\|g_t\right\|_*^2 \\
&= \frac{1}{\alpha_T} D_X^2 + \sum_{t=1}^T \frac{\alpha_t}{2}\left\|g_t\right\|_*^2,
\end{aligned}$$

where inequality (iii) follows because $\alpha_t \le \alpha_{t-1}$ and $D_X^2 \ge D_h(x^\star, x_t) \ge 0$.

(b) We have for any $a, b \ge 0$ that

$$\inf_{\alpha \ge 0} \left\{\frac{a}{\alpha} + b\alpha\right\} = 2\sqrt{ab}$$

by taking derivatives and setting to zero, as $1/\alpha$ is convex and so is $b \cdot \alpha$ (we take $\alpha = \sqrt{a/b}$). Thus

$$\inf_{\alpha \ge 0} \left\{\sum_{t=1}^T \frac{D_X^2}{\alpha} + \sum_{t=1}^T \frac{\alpha}{2}\left\|g_t\right\|_*^2\right\} = \sqrt{2D_X^2 \sum_{t=1}^T \left\|g_t\right\|_*^2}.$$

(c) We prove the result inductively. The base case is immediate: we certainly have $a_1/\sqrt{a_1} \le \sqrt{a_1} \le 2\sqrt{a_1}$. For the induction assume that $\sum_{\tau=1}^{t-1} \frac{a_\tau}{\sqrt{b_\tau}} \le 2\sqrt{b_{t-1}}$. We have

$$\sum_{\tau=1}^t \frac{a_\tau}{\sqrt{b_\tau}} \le 2\sqrt{b_{t-1}} + \frac{a_t}{\sqrt{b_t}}.$$

The first-order concavity inequality that $\phi(y) \le \phi(x) + \phi'(x)(y - x)$ for ϕ concave applies to $\sqrt{\cdot}$, guaranteeing that $\sqrt{x + \delta} \le \sqrt{x} + \frac{1}{2\sqrt{x}}\delta$. Thus we have

$$\sqrt{b_{t-1}} = \sqrt{b_t - a_t} \le \sqrt{b_t} - \frac{a_t}{2\sqrt{b_t}} \quad \text{so} \quad 2\sqrt{b_{t-1}} + \frac{a_t}{\sqrt{b_t}} \le 2\sqrt{b_t} - \frac{a_t}{2\sqrt{b_t}} + \frac{a_t}{\sqrt{b_t}} = 2\sqrt{b_t},$$

which is the inductive result we desired.

(d) Take stepsizes

$$\alpha_t = \frac{D_X}{\sqrt{\sum_{\tau=1}^t \left\|g_\tau\right\|_*^2}}.$$

Then we have

$$\mathsf{Reg}_T \le D_X \sqrt{\sum_{t=1}^T \left\|g_t\right\|_*^2} + \frac{D_X}{2} \sum_{t=1}^T \frac{\left\|g_t\right\|_*^2}{\sqrt{\sum_{\tau=1}^t \left\|g_\tau\right\|_*^2}} \le D_X \sqrt{\sum_{t=1}^T \left\|g_t\right\|_*^2} + D_X \sqrt{\sum_{t=1}^T \left\|g_t\right\|_*^2}$$

where we have applied part (c) with $a_t = \left\|g_t\right\|_*^2$. The constant $O(1)$ term is thus $O(1) \le \sqrt{2}$.

□

Question 4 (AdaGrad)**:** We investigate subgradient methods that change the metric they use throughout the iterations. In particular, we consider a sequence $H_t \in \mathbb{R}^{d\times d}$ of symmetric, diagonal, positive definite matrices, which we generate sequentially (this is AdaGrad) as follows:

i. Receive f_t and compute $g_t \in \partial f_t(x_t)$

ii. Set $G_t = \sum_{\tau=1}^{t} \operatorname{diag}(g_\tau)^2$ and $H_t = G_t^{\frac{1}{2}}$

iii. Update

$$x_{t+1} = \operatorname*{argmin}_{x\in X} \left\{ \langle g_t, x\rangle + \frac{1}{2\alpha}(x - x_t)^T H_t (x - x_t) \right\}.$$

Here $\alpha > 0$ is a fixed multiplier.

(a) Show that for any $x^\star \in X$,

$$\sum_{t=1}^{T} [f_t(x_t) - f_t(x^\star)] \le \frac{1}{2\alpha} \operatorname{tr}(H_T) \sup_{x,y\in X} \|x - y\|_\infty^2 + \sum_{t=1}^{T} \frac{\alpha}{2} \|g_t\|_{H_t^{-1}}^2$$

where $\|x\|_A^2 = x^T A x$ is the usual Mahalanobis norm

(b) Let $D_\infty = \sup_{x,y\in X} \|x - y\|_\infty$. Show that the choice $\alpha = D_\infty$ yields

$$\sum_{t=1}^{T} [f_t(x_t) - f_t(x^\star)] \le 2 \operatorname{tr}(H_T) D_\infty.$$

(c) Suppose that $X = [-1, 1]^d$ is the ℓ_∞-box in $\mathbb{R}^d$ of radius 1 and that $\|g_t\|_2 \le 1$ for all t. Give an upper bound on the regret of AdaGrad in this case. How does it compare to the regret bound one would achieve using the standard projected subgradient method?

(d) Suppose that $X = [-1, 1]^d$ as above and that instead of the fully adversarial setting, the functions f_t are drawn i.i.d. with expectation $F = \mathbb{E}[f_t]$ and that the subgradients $g_t \in \partial f_t(x_t)$ are *sparse* as follows. We have $g_t \in \{-1, 0, 1\}^d$, with coordinates $g_{t,j} \in \{-1, 0, 1\}$, and

$$\mathbb{P}(g_{t,j} \ne 0) = j^{-\beta}$$

for some $\beta \in [0, 2]$. Give an upper bound on

i. The expected regret of AdaGrad.

ii. The expected regret of the standard projected subgradient method.

In which circumstances is one better than the other?

Answer:

(a) By following the usual calculation as done in the lecture notes (see mirror descent slides), we find that the progress of a single update is

$$f_t(x_t) - f_t(x^\star) \le \frac{1}{2\alpha}\left[\|x_t - x^\star\|_{H_t}^2 - \|x_{t+1} - x^\star\|_{H_t}^2\right] + \frac{\alpha}{2}\|g_t\|_{H_t^{-1}}^2 .$$

The sum over t is then

$$\sum_{t=1}^{T}[f_t(x_t) - f_t(x^\star)] \le \frac{1}{2\alpha}\underbrace{\left[\|x_1 - x^\star\|_{H_1}^2 + \sum_{t=2}^{T}\left[\|x_t - x^\star\|_{H_t}^2 - \|x_t - x^\star\|_{H_{t-1}}^2\right]\right]}_{(*)} + \frac{\alpha}{2}\sum_{t=1}^{T}\|g_t\|_{H_t^{-1}}^2$$

where $(*)$ is

$$\begin{aligned}(*) &= (x_1 - x^\star)^T H_1 (x_1 - x^\star) + \sum_{t=2}^{T}(x_t - x^\star)^T (H_t - H_{t-1})(x_t - x^\star) \\ &\le D_\infty^2 \operatorname{tr}(H_1) + \sum_{t=2}^{T} D_\infty^2(\operatorname{tr}(H_t) - \operatorname{tr}(H_{t-1})) = D_\infty^2 \operatorname{tr}(H_T),\end{aligned}$$

since each H_t is diagonal with elements greater than those of H_{t-1}. (Here we have denoted $D_\infty = \sup_{x,y\in X}\|x - y\|_\infty$.) This produces the desired bound

$$\sum_{t=1}^{T}[f_t(x_t) - f_t(x^\star)] \le \frac{1}{2\alpha}D_\infty^2 \operatorname{tr}(H_T) + \frac{\alpha}{2}\sum_{t=1}^{T}\|g_t\|_{H_{t-1}}^2 .$$

(b) By definition of the Mahalanobis norm,

$$\sum_{t=1}^{T}\|g_t\|_{H_t^{-1}}^2 = \sum_{t=1}^{T}\sum_{j=1}^{d}\frac{g_{t,j}^2}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,j}^2}} \le 2\sum_{j=1}^{d}\sqrt{\sum_{t=1}^{T} g_{t,j}^2} = 2\sum_{j=1}^{d} H_{T,j} = 2\operatorname{tr}(H_T),$$

where we have reversed the sums and applied the result of Problem 3(c). Now if $\alpha = D_\infty$, the bound from part (a) becomes

$$\sum_{t=1}^{T}[f_t(x_t) - f_t(x^\star)] \le \frac{D_\alpha}{2}\operatorname{tr}(H_T) + \frac{D_\alpha}{2}\sum_{t=1}^{T}\|g_t\|_{H_t^{-1}}^2 \le 2D_\infty \operatorname{tr}(H_T).$$

(c) In this case $D_\infty = 2$, and $\|g_t\|_2 \le 1$ implies

$$\operatorname{tr}(H_T) = \sum_{j=1}^{d}\sqrt{\sum_{t=1}^{T} g_{t,j}^2} \le \sqrt{d\sum_{t=1}^{T}\sum_{j=1}^{d} g_{t,j}^2} \le \sqrt{dT}.$$

Therefore, the regret bound from part (b) is $4\sqrt{dT}$.

This has the same $\sqrt{dT}$ dependence as the standard projected subgradient method.

(d) The expected regret is bounded by

$$\mathbb{E}\left[\sum_{i=1}^{T}[f_t(x_t) - f_t(x^\star)]\right] \le 2D_\infty \mathbb{E}[\mathrm{tr}(H_T)] \le 2D_\infty \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} \mathbb{E}[g_{t,j}^2]} = 2\sqrt{T}\sum_{j=1}^{d} j^{-\beta/2} \le C\sqrt{T} d^{1-\beta/2},$$

for some constant C.

From class, the expected regret of the standard projected subgradient method is bounded above by $D_X\sqrt{T}M$ where where $D_X = \sup_{x,y\in X} \|x - y\|_2 = 2\sqrt{d}$, and

$$M^2 = \sum_{j=1}^{d} \mathbb{P}(g_{t,j} \neq 0) = \sum_{j=1}^{d} j^{-\beta} \asymp \begin{cases} 1 & \text{if } \beta > 1 \\ \log d & \text{if } \beta = 1 \\ d^{1-\beta} & \text{if } \beta < 1. \end{cases}$$

Ignoring the logarithmic case for simplicity, we have $M \asymp d^{[1-\beta]_+/2}$, so that

$$\mathbb{E}\left[\sum_{t=1}^{T}[f_t(x_t) - f_t(x^\star)]\right] \le D_X\sqrt{T}M \lesssim \sqrt{T} d^{\frac{1}{2}+[1-\beta]_+/2}.$$

Evidently, this is worse than AdaGrad's $d^{1-\beta/2}$ dependence for all $\beta \ge 1$.

□

Question 5 (Strongly convex regret)**:** Assume that we have an online convex optimization problem where each $f_t : X \to \mathbb{R}$ is λ-strongly convex, meaning

$$f_t(y) \ge f_t(x) + \langle g_t, y - x\rangle + \frac{\lambda}{2}\|x - y\|_2^2 \quad \text{for } g_t \in \partial f(x) \text{ and } x, y \in X.$$

Assume that each f_t is also M-Lipschitz, so that $\|g\|_2 \le M$ for all $g \in \partial f(x)$, $x \in X$. Prove that for the usual projected gradient algorithm,

$$x_{t+1} = \pi_X(x_t - \alpha_t g_t),$$

where $g_t \in \partial f_t(x_t)$ and we choose the stepsize $\alpha_t = \frac{1}{\lambda t}$, we have

$$\mathsf{Reg}_T \le \frac{M^2}{2\lambda}\log(T+1).$$

Answer: We follow our usual proof for these types of results—we expand the error $\frac{1}{2}\|x_{t+1} - x\|_2^2$. We have for any $x \in X$ that

$$\begin{aligned}
\frac{1}{2}\|x_{t+1} - x\|_2^2 &\le \frac{1}{2}\|x_t - \alpha_t g_t - x\|_2^2 \\
&= \frac{1}{2}\|x_t - x\|_2^2 - \alpha_t\langle g_t, x_t - x\rangle + \frac{\alpha_t^2}{2}\|g_t\|_2^2 \\
&\le \frac{1}{2}\|x_t - x\|_2^2 - \alpha_t\left[f_t(x_t) - f_t(x) + \frac{\lambda}{2}\|x_t - x\|_2^2\right] + \frac{\alpha_t^2}{2}\|g_t\|_2^2,
\end{aligned}$$

where the final step used the definition of strong convexity. Rearranging and dividing by α_t yields

$$f_t(x_t) - f_t(x) \le \frac{1}{2\alpha_t}\left\|x_t - x\right\|_2^2 - \frac{1}{2\alpha_t}\left\|x_{t+1} - x\right\|_2^2 - \frac{\lambda}{2}\left\|x_t - x\right\|_2^2 + \frac{\alpha_t}{2}\left\|g_t\right\|_2^2.$$

Noting that $\left\|g_t\right\|_2 \le M$ by assumption, we sum the preceding inequality from $t = 1$ to T to obtain

$$\begin{aligned}
\sum_{t=1}^{T}\left[f_t(x_t) - f_t(x)\right] &\le \sum_{t=2}^{T}\left(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}\right)\left\|x_t - x\right\|_2^2 + \frac{1}{2\alpha_1}\left\|x_1 - x\right\|_2^2 - \frac{1}{2\alpha_T}\left\|x_{T+1} - x\right\|_2^2 \\
&\quad - \sum_{t=1}^{T}\frac{\lambda}{2}\left\|x_t - x\right\|_2^2 + \sum_{t=1}^{T}\frac{\alpha_t}{2}M^2 \\
&= \sum_{t=1}^{T}\frac{\lambda}{2}\left\|x_t - x\right\|_2^2 - \sum_{t=1}^{T}\frac{\lambda}{2}\left\|x_t - x\right\|_2^2 - \frac{1}{2\alpha_T}\left\|x_{T+1} - x\right\|_2^2 + \sum_{t=1}^{T}\frac{\alpha_t}{2}M^2 \\
&\le \sum_{t=1}^{T}\frac{\alpha_t}{2}M^2,
\end{aligned}$$

where the equality uses that $\frac{1}{\alpha_t} = t\lambda$. Noting that $\sum_{t=1}^{T}\frac{1}{t} \le \int_1^{T+1}\frac{1}{t}dt = \log(T+1)$ gives the result. □

Question 6 (Low regret algorithms prove von-Neumann's Minimax Theorem)**:** A minor extension of the von-Neumann minimax theorem is as follows. Let $A \in \mathbb{R}^{m\times n}$ be an arbitrary matrix, and let $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ be arbitrary convex compact sets. Then

$$\inf_{x\in X}\sup_{y\in Y} x^T A y = \sup_{y\in Y}\inf_{x\in X} x^T A y. \tag{3}$$

In fact, we can say more: there exists a saddle point $x^\star, y^\star$ such that

$$\inf_{x\in X} x^T A y^\star = x^{\star T} A y^\star = \sup_{y\in Y} x^{\star T} A y.$$

In this question, we show how *online learning* gives a proof of the von-Neumann minimax theorem. Throughout this question, with no loss of generality, we assume that $\left|\!\left|\!\left|A\right|\!\right|\!\right|_{\rm op} \le 1$ and $\left\|x - x'\right\|_2 \le 1$, $\left\|y - y'\right\|_2 \le 1$ for all $x, x' \in X$ and $y, y' \in Y$.

(a) Show the "easy" direction

$$\sup_{y\in Y}\inf_{x\in X} x^T A y \le \inf_{x\in X}\sup_{y\in Y} x^T A y.$$

Consider the following so-called "best response" game: beginning from an arbitrary $x_1 \in X$, at each iteration $t = 1, 2, \ldots$, we play

$$y_t = \operatorname*{argmax}_{y\in Y}\left\{x_t^T A y\right\}$$

and update

$$x_{t+1} = \operatorname*{argmin}_{x\in X}\left\{x^T A y_t + \frac{1}{2\alpha}\left\|x - x_t\right\|_2^2\right\},$$

or $x_{t+1} = \pi_X(x_t - \alpha A y_t)$, the projection of $x_t - \alpha A y_t$ onto X.

(b) Defining $f_t(x) = x^T A y_t$, give an upper bound on

$$\mathsf{Reg}_T := \sup_{x \in X} \sum_{t=1}^{T} [f_t(x_t) - f_t(x)]$$

that, for appropriate choice of α, satisfies $\mathsf{Reg}_T \le \sqrt{T}$.

(c) Show that for $\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\overline{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$, we have

$$\sup_{y \in Y} \overline{x}_T^T A y \le \inf_{x \in X} x^T A \overline{y}_T + \frac{1}{\sqrt{T}}.$$

Show that this gives von-Neumann's result (3). (It turns out that by moving to subsequences if necessary, this argument also shows that $\overline{x}_T \to x^\star$ and $\overline{y}_T \to y^\star$ as $T \to \infty$.)

Answer:

(a) For any fixed x and y, it is clear that

$$x^T A y \le \sup_{y' \in Y} x^T A y'.$$

Taking the infimum in x over both sides gives

$$\inf_{x \in X} x^T A y \le \inf_{x \in X} \sup_{y' \in Y} x^T A y'.$$

But then taking an infimum over y on the left gives the result.

(b) We have that $g_t := A y_t = \nabla f_t(x_t)$, so that letting $x^\star \in \operatorname{argmin}_{x \in X} \sum_{t=1}^{T} f_t(x)$ (which exists because f_t are convex and X is compact), we have

$$\sum_{t=1}^{T} [f_t(x_t) - f_t(x^\star)] \le \sum_{t=1}^{T} \langle g_t, x_t - x^\star \rangle .$$

But this is the exact upper bound that we have seen in class, so using the lecture notes, we find that the projected gradient update yields

$$\sum_{t=1}^{T} \langle g_t, x_t - x^\star \rangle \le \frac{\|x_1 - x^\star\|_2^2}{2\alpha} + \frac{\alpha}{2} \sum_{t=1}^{T} \|g_t\|_2^2 .$$

Using that $\|g_t\|_2 = \|A y_t\|_2 \le |\!|\!| A |\!|\!|_{\mathrm{op}} \|y_t\|_2 \le |\!|\!| A |\!|\!|_{\mathrm{op}} \le 1$, we have

$$\mathsf{Reg}_T \le \frac{1}{2\alpha} + \frac{\alpha}{2} T.$$

Choose $\alpha = 1/\sqrt{T}$.

(c) Letting $\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\overline{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$, we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} f_t(x_t) = \frac{1}{T}\sum_{t=1}^{T} x_t^T A y_t &\overset{(i)}{\le} \inf_{x \in X} \left\{ \frac{1}{T}\sum_{t=1}^{T} x^T A y_t \right\} + \frac{1}{\sqrt{T}} \\ &= \inf_{x \in X} x^T A \overline{y}_T + \frac{1}{\sqrt{T}} \le \sup_{y \in Y} \inf_{x \in X} x^T A y + \frac{1}{\sqrt{T}}, \end{aligned}$$

where inequality (i) is part (b). But then $f_t(x_t) = \sup_{y \in Y} x_t^T A y$ by the choice of y_t, so

$$\frac{1}{T} \sum_{t=1}^{T} \sup_{y \in Y} \{x_t A y\} \leq \inf_{x \in X} x^T A \overline{y}_T + \frac{1}{\sqrt{T}}.$$

Noting that the average of suprema is larger than the supremum of the average (concavity), we have

$$\inf_{x \in X} \sup_{y \in X} x^T A y \leq \sup_{y \in Y} \overline{x}_T^T A y \leq \frac{1}{T} \sum_{t=1}^{T} \sup_{y \in Y} x_t^T A y \leq \inf_{x \in X} x^T A \overline{y}_T + \frac{1}{\sqrt{T}} \leq \sup_{y \in Y} \inf_{x \in X} x^T A y + \frac{1}{\sqrt{T}}.$$

As T is arbitrary, we have the result we desire.

□