COMS 4773 Spring 2024 HW 2 (due Mar. 1 at noon)

Please read the handout on *McDiarmid's inequality*, posted on the course website.

Theorem 1 (McDiarmid's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, where X_i has range $\mathcal{X}_i$. Let $f \colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_n)$-*bounded differences property*: for every $i = 1, \ldots, n$ and every $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ that differ only in the i-th coordinate ($x_j = x'_j$ for all $j \neq i$), we have*

$$|f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n)| \leq c_i.$$

For any $t > 0$,

$$\Pr(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Problem 1. First, check for yourself that Hoeffding's inequality is a simple corollary of McDiarmid's inequality.

(a) Suppose $\mathcal{F}$ is a (possibly infinite) collection of real-valued functions on $\mathcal{X}$, each with range $[a, b]$, μ is a probability distribution over $\mathcal{X}$, and S is an iid sample from μ of size n. For any $f \in \mathcal{F}$, let $\mu(f) = \mathbb{E}_{X \sim \mu}[f(X)]$ and $\mu_S(f) = \frac{1}{n} \sum_{x \in S} f(x)$. Use McDiarmid's inequality to prove the following: for any $t > 0$,

$$\Pr\left(\max_{f \in \mathcal{F}} |\mu(f) - \mu_S(f)| - \mathbb{E}\left[\max_{f \in \mathcal{F}} |\mu(f) - \mu_S(f)|\right] \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

(b) Suppose $p = (p_1, \ldots, p_k) \in \Delta^{k-1}$ is a probability distribution over $\{1, \ldots, k\}$, and $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_k)$ is the empirical probability distribution based on an iid sample from p of size n, i.e.,

$$\hat{p}_i = \frac{\text{number of times } i \text{ appears in the iid sample}}{n}.$$

Prove the following: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|p - \hat{p}\|_2 \leq \sqrt{\frac{1 - \|p\|_2^2}{n}} + \sqrt{\frac{\ln(1/\delta)}{n}}.$$

Hint: McDiarmid's inequality is useful for part of this problem.

Problem 2. Suppose $X_1, \ldots, X_n$ are 1-subgaussian mean-zero random variables (not necessarily independent nor identically distributed), and let

$$Z := \max_{i=1,\ldots,n} X_i.$$

In this problem, you will prove a bound on $\mathbb{E}[Z]$ two (or three) ways. (Throughout this problem, assume $n > 1$ so $\log(n) > 0$. This will simplify the form of the bounds.)

(a) Prove the following: for any $\lambda > 0$,

$$\mathbb{E}[Z] \leq \frac{1}{\lambda} K_Z(\lambda),$$

where $K_Z(\lambda) = \ln \mathbb{E}[\exp(\lambda Z)]$ is the log moment generating function for Z.

Hint: Use Jensen's inequality.

(b) Use the result from Part (a) to prove the following: for some absolute constant $C > 0$,

$$\mathbb{E}[Z] \leq C\sqrt{\log(n)}.$$

Hint: For any real numbers $a_1, \ldots, a_n$, we have $e^{\max_{i \in [n]} a_i} = \max_{i \in [n]} e^{a_i} \leq \sum_{i=1}^n e^{a_i}$.

Now we start the second way to prove the same bound on $\mathbb{E}[Z]$.

(c) Prove the following: for any $t > 0$,

$$\Pr(Z \geq t) \leq n \cdot e^{-t^2/2}.$$

Hint: This one is easy; not a trick question.

(d) Use the result from Part (c) and the fact that $\mathbb{E}[Z] \leq \int_0^\infty \Pr(Z \geq t)\,\mathrm{d}t$ to prove the following: for some absolute constant $C > 0$,

$$\mathbb{E}[Z] \leq C\sqrt{\log(n)}.$$

Hint: Break the integral into two parts, $[0, w)$ and $[w, \infty)$, for some judicious choice of $w > 0$; use a "trivial" bound for the first part, and use a bound that takes advantage of the lower integral limit for the second part.

Here is a generalization.

(e) (Optional.) Suppose $X_1, X_2, \ldots$ is an infinite sequence of 1-subgaussian mean-zero random variables (not necessarily independent nor identically distributed), and let

$$Y := \max_{i=1,2,\ldots} \frac{X_i}{\sqrt{1 + \ln(i)}}.$$

Prove the following: for some absolute constant $C > 0$,

$$\mathbb{E}[Y] \leq C.$$

Problem 3. Suppose μ is a probability distribution over $\mathcal{X} \times \{0,1\}$. Consider the following online prediction problem that unfolds over the course of T rounds. In each round $t = 1, \dots, T$:

1. First, Nature independently draws a random example (X_t, Y_t) from μ, and reveals X_t to the learner (but Y_t is kept hidden).
2. Next, the learner makes a prediction $\hat{Y}_t$ of Y_t.
3. Finally, Nature reveals the label Y_t to the learner.

Let M_T be the number of mistakes made by the learner in all T rounds:

$$M_T = \sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\}.$$

Let $\mathcal{H}$ be a finite hypothesis class of functions mapping $\mathcal{X}$ to $\{0,1\}$,[1] and for each $h \in \mathcal{H}$, let $M_{T,h}$ be the number of mistakes made by hypothesis h in all T rounds:

$$M_{T,h} = \sum_{t=1}^{T} \mathbb{1}\{h(X_t) \neq Y_t\}.$$

(a) Explain how to use RANDOMIZED WEIGHTED MAJORITY (with a suitable choice of the hyperparameter) for this problem to guarantee

$$\mathbb{E}[M_T - M_{T,h}] \leq O\left(\sqrt{T \log|\mathcal{H}|}\right) \quad \text{for all } h \in \mathcal{H}.$$

(b) Consider the following algorithm for this problem. Let $\hat{h}_1 \in \mathcal{H}$ be any arbitrary hypothesis from $\mathcal{H}$; in round $t > 1$, let

$$\hat{h}_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbb{1}\{h(X_s) \neq Y_s\}$$

be a hypothesis that makes the fewest mistakes in all previous rounds; set $\hat{Y}_t := \hat{h}_t(X_t)$. For this algorithm, prove the following:

$$\mathbb{E}[M_T - M_{T,h}] \leq O\left(\sqrt{T \log|\mathcal{H}|}\right) \quad \text{for all } h \in \mathcal{H}.$$

[1] The notation $\mathcal{Y}^{\mathcal{X}}$ is used to denote the set of all possible functions from $\mathcal{X}$ to $\mathcal{Y}$, so $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$.

Problem 4. Suppose μ is a probability distribution on $\mathcal{X} \times \{0,1\}$, and $(\mathcal{H}_k)_{k\in\mathbb{N}}$ is an infinite sequence of finite hypothesis classes on $\mathcal{X}$, where $2 \le |\mathcal{H}_1| < |\mathcal{H}_2| < \cdots$. (A typical setup is one where the classes are nested: $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$.) Let S denote an iid sample from μ of size n, and define

$$\begin{aligned} \operatorname{err}(h) &= \Pr_{(X,Y)\sim\mu}(h(X) \neq Y), \\ \widehat{\operatorname{err}}(h) &= \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{h(x) \neq y\}. \end{aligned}$$

(a) Prove that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|\operatorname{err}(h) - \widehat{\operatorname{err}}(h)| \le \operatorname{bound}(k,n,\delta) \quad \text{for all } k \ge 1 \text{ and all } h \in \mathcal{H}_k$$

where

$$\operatorname{bound}(k,n,\delta) := C\sqrt{\frac{\log|\mathcal{H}_k| + \log(k) + \log(1/\delta)}{n}}$$

for some absolute constant $C > 0$.

Hint: Use Hoeffding's inequality and union bound, together with the fact

$$\sum_{k=1}^{\infty} \frac{1}{k^2+k} = \sum_{k=1}^{\infty} \frac{1}{k} - \frac{1}{k+1} = 1.$$

(b) Consider the following strategy for choosing $\hat{h} \in \bigcup_{k\ge1} \mathcal{H}_k$. Define

$$\hat{h}_k := \operatorname*{arg\,min}_{h\in\mathcal{H}_k} \widehat{\operatorname{err}}(h) \quad \text{for each } k \ge 1,$$

assuming ties are broken in some way. Choose

$$\hat{k} := \operatorname*{arg\,min}_{k\ge1} \min\{1, \widehat{\operatorname{err}}(\hat{h}_k) + \operatorname{bound}(k,n,\delta)\},$$

assuming ties are broken in favor of the smaller k, and set $\hat{h} := \hat{h}_{\hat{k}}$.

The strategy above is simple to write down, but the "$\arg\min_{k\ge1}$" should give some pause. Briefly explain how the strategy can be executed in finite time. (Assume, for each k, that you have an algorithm for computing both $\hat{h}_k$ and $\operatorname{bound}(k,n,\delta)$.)

(c) (Continuing from Part (b).) Prove the following: for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\operatorname{err}(\hat{h}) \le \min_{k\ge1} \operatorname{err}(h_k^*) + 2\operatorname{bound}(k,n,\delta)$$

where $h_k^* := \arg\min_{h\in\mathcal{H}_k} \operatorname{err}(h)$ for each $k \ge 1$.

Problem 5. Recall the following online convex optimization problem. In round $t = 1, 2, \ldots$:

1. The learner chooses $x_t \in \mathbb{R}^n$.
2. Nature chooses (differentiable) convex function $f_t \colon \mathbb{R}^n \to \mathbb{R}$ and reveals $\nabla f_t(x_t)$ to the learner.
3. The learner incurs loss $f_t(x_t)$.

In this problem, you will analyze a variant of the online gradient descent algorithm that chooses the x_t's as follows:

$$x_t := \arg\min_{x \in \mathbb{R}^n} \sum_{s=1}^{t-1} \langle \nabla f_s(x_s), x \rangle + \langle \hat{\ell}_t, x \rangle + \frac{\|x\|_2^2}{2\eta}, \tag{1}$$

where $\hat{\ell}_1, \hat{\ell}_2, \ldots$ is some arbitrary sequence of vectors in $\mathbb{R}^n$ with $\hat{\ell}_1 = 0$. (When $t = 1$, the sum is empty and $\hat{\ell}_1 = 0$, so $x_1 = \arg\min_{x \in \mathbb{R}^n} \|x\|_2^2/(2\eta) = 0$.)

The idea of the $\hat{\ell}_t$'s is that they are "guesses" for the actual gradients $\ell_t = \nabla f_t(x_t)$. In round t, the gradient ℓ_t is not available to the learner, so the learner has to make do with $\hat{\ell}_t$. Don't worry about where these guesses might come from for now. This algorithm has the following guarantee: for any T and any $x \in \mathbb{R}^n$,

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x) \le \eta \sum_{t=1}^{T} \|\nabla f_t(x_t) - \hat{\ell}_t\|_2^2 + \frac{\|x\|_2^2}{2\eta}. \tag{2}$$

Here are some possible interpretations of (2).

- If $\hat{\ell}_t = 0$ for all t, then the algorithm is the same as the usual online gradient descent, and the right-hand side of (2) is the same guarantee we had before.
- If $\hat{\ell}_t = \ell_t$ for all t (i.e., the "guesses" are perfect!), then the right-hand side of (2) is just $\|x\|^2/(2\eta)$, which does not grow with the number of rounds T at all!
- If $\hat{\ell}_t = \ell_{t-1}$, then the right-hand side of (2) is

  $$\eta \sum_{t=1}^{T} \|\nabla f_t(x_t) - \nabla f_{t-1}(x_{t-1})\|_2^2 + \frac{\|x\|_2^2}{2\eta},$$

  which may be small if the gradients don't change very much from round to round.

Your task is to prove of the guarantee in (2) in two steps.

(a) Prove the following lemma.

Lemma 1. *Let $x_1, x_2, \ldots$ be the choices of the online gradient descent variant from (1). Define another sequence $x_1^{\text{ogd}}, x_2^{\text{ogd}}, \ldots$ by*

$$x_t^{\text{ogd}} := \arg\min_{x \in \mathbb{R}^n} \sum_{s=1}^{t-1} \langle \nabla f_s(x_s), x \rangle + \frac{1}{2\eta} \|x\|_2^2. \tag{3}$$

For any T and any $x \in \mathbb{R}^n$,

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, x_t - x_{t+1}^{\text{ogd}} \rangle + \langle \ell_t, x_{t+1}^{\text{ogd}} \rangle \leq \sum_{t=1}^{T} \langle \ell_t, x \rangle + \frac{\|x\|_2^2}{2\eta}.$$

Hint: Use induction on T. The base case ($T = 1$) uses the fact that $\hat{\ell}_1 = 0$. For the inductive step, use the inductive hypothesis with a careful choice of x. You should only have to use the "optimality" properties guaranteed by the definitions of x_t and x_t^{ogd}.

(b) Armed with Lemma 1 from Part (a), prove the guarantee in (2).

Hint: It may be helpful to obtain explicit expressions for x_t and x_{t+1}^{ogd} (as defined in (1) and (3)). The decomposition

$$\langle \nabla f_t(x_t), x_t - x \rangle = \langle \nabla f_t(x_t) - \hat{\ell}_t, x_t - x_{t+1}^{\text{ogd}} \rangle + \langle \hat{\ell}_t, x_t - x_{t+1}^{\text{ogd}} \rangle + \langle \nabla f_t(x_t), x_{t+1}^{\text{ogd}} - x \rangle$$

may also be helpful.