Problem Set 2

Stats 311/EE 377

Due: Thursday, February 4 in class

Question 2.1 (A discrete isoperimetric inequality)**:** Let $A \subset \mathbb{Z}^d$ be a finite subset of the d-dimensional integers. Let the projection mapping $\pi_j : \mathbb{Z}^d \to \mathbb{Z}^{d-1}$ be defined by

$$\pi_j(z_1, \ldots, z_d) = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_d)$$

so that we "project out" the jth coordinate, and define the projected sets.

$$\begin{aligned} A_j = \pi_j(A) &= \{\pi_j(z) : z \in A\} \\ &= \left\{z \in \mathbb{Z}^{d-1} : \text{ there exists } z_\star \in \mathbb{Z} \text{ such that } (z_1, z_2, \ldots, z_{j-1}, z_\star, z_j, \ldots, z_{d-1}) \in A\right\}. \end{aligned}$$

Prove the Loomis-Whitney inequality, that is, that

$$\mathrm{card}(A) \le \left(\prod_{j=1}^{d} \mathrm{card}(A_j)\right)^{\frac{1}{d-1}}.$$

Question 2.2 (Optimal algorithms for memory access)**:** In a modern CPU, memory is organized in a hierarchy, so that data upon which computations are being actively performed lies in a very small memory close to the logic units of the processor for which access is extraordinarily fast, while data not being actively used lies in slower memory slightly farther from the processor. (Modern processor memory is generally organized into the registers—a small number of 4- or 8-byte memory locations on the processor—and several increasing-sized levels of cache, which contain small amounts of data and increasing access times, and RAM (random access memory).) Moving data—communicating—between levels of the memory hierarchy is both power intensive and slow relative to computation on the data itself, so that in many algorithms the bulk of the time of the algorithm is in moving data from one place to another. Thus, developing very fast algorithms for numerical (and other) tasks on modern computers requires careful tracking of memory access, and control of these quantities can often yield orders of magnitude speed improvements in execution. In this problem, we prove a lower bound on the number of communication steps that certain numerical-type methods must perform, giving a concrete (attainable) inequality that allows one to certify optimality of *specific* algorithms.

We consider matrix multiplication, as it is a proxy for a class of cubic algorithms that are well behaved. Let $A, B \in \mathbb{R}^{n \times n}$ be matrices, and assume we wish to compute $C = AB$, via the simple algorithm that for all i, j sets

$$C_{ij} = \sum_{l=1}^{n} A_{il} B_{lj}.$$

Computationally, this forces us to repeatedly execute operations of the form

$$\mathsf{Mem}(C_{ij}) = F(\mathsf{Mem}(A_{il}), \mathsf{Mem}(B_{lj}), \mathsf{Mem}(C_{ij})),$$

where F is some function—that may depend on i, j, l—and $\mathsf{Mem}(\cdot)$ indicates that we access the memory associated with the argument. (In our case, we have $C_{ij} = C_{ij} + A_{il} \cdot B_{lj}$.) We assume that executing F requires that $\mathsf{Mem}(A_{il})$, $\mathsf{Mem}(B_{lj})$, and $\mathsf{Mem}(C_{ij})$ belong to fast memory, and that each are distinct (stored in a separate place in slow and fast memory). We assume that the order of the computations does *not* matter, so we may re-order them in any way to improve memory access. We call $\mathsf{Mem}(A_{il})$ (respectively B or C) and *operand* in our computation. We let M denote the size of fast/local memory, and we would like to lower bound the number of times we must communicate an operand into or out of the fast local memory when all we may do is re-order the computation being executed. We let N_{Store} denote the number of times we write something from fast memory out to slow memory and let N_{Load} the number of times we load something from slow memory to fast memory. Let N be the total number of operations we execute (for simple matrix multiplication, we have $N = n^3$, though with sparse matrices, this can be smaller).

We analyze the procedure by breaking the computation into a number of segments, where each segment contains precisely M load or store (communication-causing) instructions.

(a) Let N_{seg} be an upper bound on the number of evaluations with the function $F(\cdot)$ in any given segment (you will upper bound this in a later part of the problem). Justify that

$$N_{\mathsf{Store}} + N_{\mathsf{Load}} \geq M \left\lfloor N / N_{\mathsf{seg}} \right\rfloor.$$

(b) Within a segment, all operands involved must be in fast memory at least once to be computed with. Assume that memory locations $\mathsf{Mem}(A_{il})$, $\mathsf{Mem}(B_{lj})$, and $\mathsf{Mem}(C_{ij})$ do not overlap. For any operand involved in a memory operation in one of the segments, the operand (1) was already in fast memory at the beginning of the segment, (2) was read from slow memory, (3) is still in fast memory at the end of the segment, or (4) is written to slow memory at the end of the segment. (There are also operands potentially created during execution that are simply discarded; we do not bound those.) Justify the following: within a segment, for each type of operand $\mathsf{Mem}(A_{ij})$, $\mathsf{Mem}(B_{ij})$, or $\mathsf{Mem}(C_{ij})$, there are at most $c \cdot M$ such operands (i.e. there are at most cM operands of type $\mathsf{Mem}(A_{ij})$, independent of the others, and so on), where c is a numerical constant. What value of c can you attain?

(c) Using the result of question 2.1, argue that $N_{\mathsf{seg}} \leq c' \sqrt{M^3}$ for a numerical constant c'. What value of c' do you get?

(d) Using the result of part (c), argue that the number of loads and stores satisfies

$$N_{\mathsf{Store}} + N_{\mathsf{Load}} \geq c'' \frac{N}{\sqrt{M}} - M$$

for a numerical constant c''. What is your constant?

Question 2.3: Let $\mathcal{F}$ be a collection of functions or parameters, let $\mathcal{Z}$ be a space in which data lies, $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. In this problem, we give a so-called PAC-Bayes generalization bound, in the sense that it relies on the difference between a prior and posterior distribution (PAC

stands for "probably approximately correct"). We assume that for each $f \in \mathcal{F}$, the random variable $\ell(f, Z)$ is σ^2-sub-Gaussian, that is, when $Z \sim P$, we have

$$\mathbb{E}\left[\exp(\lambda(\ell(f, Z) - \mathbb{E}[\ell(f, Z)]))\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Let $R(f) = \mathbb{E}[\ell(f, Z)]$ denote the risk of a particular prediction function f, and let $\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i)$ denote the empirical risk for $Z_i \stackrel{\text{i.i.d.}}{\sim} P$. For any distribution Q on the set $\mathcal{F}$, we also define the expected risk (averaged over Q) as

$$R(Q) := \int R(f)dQ(f) = \mathbb{E}_{f\sim Q}[R(f)] \quad \text{and} \quad \widehat{R}_n(Q) := \int \widehat{R}_n(f)dQ(f) = \mathbb{E}_{f\sim Q}[\widehat{R}_n(f)].$$

We will prove that for any distribution Π on the functions in $\mathcal{F}$ (the *prior*) with probability at least $1 - \delta$ over the draw of the sample $Z_{1:n}$, that

$$\left|\widehat{R}_n(Q) - R(Q)\right| \leq \sqrt{e\sigma^2 \frac{\log\frac{2}{\delta} + D_{\rm kl}\left(Q\|\Pi\right)}{n}} \tag{1}$$

holds for *all* distributions Q (the *posterior*).

(a) Let $\Delta_n(f) = \widehat{R}_n(f) - R(f)$ for shorthand. Argue that for any fixed f,

$$\mathbb{E}_P\left[\exp\left(\frac{\Delta_n(f)^2 n}{e\sigma^2}\right)\right] < 2,$$

where the expectation $\mathbb{E}_P$ is taken over the sample $Z_{1:n}$.

(b) Let the event $\mathcal{E}$ (a function of the random sample $Z_{1:n}$) be defined by

$$\mathcal{E} := \left\{\mathbb{E}_\Pi\left[\exp\left(\frac{\Delta_n(f)^2 n}{e\sigma^2}\right) \mid Z_{1:n}\right] \geq \frac{2}{\delta}\right\},$$

where the expectation is taken over $f \sim \Pi$ and $Z_{1:n}$ are held fixed. Argue that $\mathbb{P}(\mathcal{E}) < \delta$.

(c) Argue that if there exists any distribution Q on the set $\mathcal{F}$ such that

$$\mathbb{E}_Q\left[\frac{\Delta_n(f)^2 n}{e\sigma^2} \mid Z_{1:n}\right] - D_{\rm kl}\left(Q\|\Pi\right) \geq \log\frac{2}{\delta},$$

then the event $\mathcal{E}$ occurs.

(d) Show how part (c) yields inequality (1).

(e) Suppose that the class $\mathcal{F}$ is finite, and that $Z = (X, Y)$ consists of pairs $X \in \mathcal{X}$ and $Y \in \{-1, 1\}$, and that $\ell(f, Z) = 1_{(f(X)Y\leq 0)}$. Show how inequality (1) gives

$$\max_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \leq \sqrt{e\frac{\log\frac{2}{\delta} + \log|\mathcal{F}|}{4n}}.$$

(f) Suppose that the set $\mathcal{F}$ consists of functions of the form

$$f_\theta(x) = \theta^\top x$$

indexed by $\theta \in \Theta \subset \mathbb{R}^d$, and that the data pairs $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ (the standard classification problem setup in machine learning and statistics). Let the loss $\ell(f_\theta, (x, y)) = \left[1 - y\theta^\top x\right]_+$ be the hinge loss. Let Θ be the ℓ_2-ball of radius 1 in $\mathbb{R}^d$ and assume the data x satisfy $\|x\|_2 \le r$ where $r > 1$. Show that with probability at least $1 - \delta$ over the draw of the sample $Z_{1:n}$,

$$\sup_{\theta \in \Theta} \left| \widehat{R}_n(f_\theta) - R(f_\theta) \right| \le r \frac{\epsilon}{\sqrt{d}} + \sqrt{e} r \sqrt{\frac{\log \frac{2}{\delta} + d \log(1 + \frac{1}{\epsilon})}{n}}$$

for all $\epsilon \in (0, 1]$. (Hint: consider Q and Π uniform on appropriate ℓ_2-balls.)

Question 2.4: In this question, we show how to use Bernstein-type (sub-exponential) inequalities to give sharp convergence guarantees. Recall (Example 3.13, Corollary 3.17, and inequality (3.1.7) in the notes) that if X_i are independent bounded random variables with $|X_i - \mathbb{E}[X]| \le b$ for all i and $\mathrm{Var}(X_i) \le \sigma^2$, then

$$\max \left\{ \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \ge \mathbb{E}[X] + t \right), \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \le \mathbb{E}[X] - t \right) \right\} \le \exp\left(-\frac{1}{2} \min\left\{ \frac{5}{6} \frac{nt^2}{\sigma^2}, \frac{nt}{2b} \right\} \right).$$

We consider minimization of loss functions ℓ over finite function classes $\mathcal{F}$ with $\ell \in [0, 1]$, so that if $R(f) = \mathbb{E}[\ell(f, Z)]$ then $|\ell(f, Z) - R(f)| \le 1$. Throughout this question, we let

$$R^\star = \min_{f \in \mathcal{F}} R(f) \quad \text{and} \quad f^\star \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f).$$

We will show that, roughly, a procedure based on picking an empirical risk minimizer is unlikely to choose a function $f \in \mathcal{F}$ with bad performance, so that we obtain faster concentration guarantees.

(a) Argue that for any $f \in \mathcal{F}$

$$\mathbb{P}\left(\widehat{R}_n(f) \ge R(f) + t \right) \vee \mathbb{P}\left(\widehat{R}_n(f) \le R(f) - t \right) \le \exp\left(-\frac{1}{2} \min\left\{ \frac{5}{6} \frac{nt^2}{R(f)(1 - R(f))}, \frac{nt}{2} \right\} \right).$$

(b) Define the set of "bad" prediction functions $\mathcal{F}_{\epsilon\,\mathrm{bad}} := \{f \in \mathcal{F} : R(f) \ge R^\star + \epsilon\}$. Show that for any fixed $\epsilon \ge 0$ and any $f \in \mathcal{F}_{2\epsilon\,\mathrm{bad}}$, we have

$$\mathbb{P}\left(\widehat{R}_n(f) \le R^\star + \epsilon \right) \le \exp\left(-\frac{1}{2} \min\left\{ \frac{5}{6} \frac{n\epsilon^2}{R^\star(1 - R^\star) + 2\epsilon(1 - 2\epsilon)}, \frac{n\epsilon}{2} \right\} \right).$$

(c) Let $\widehat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f)$ denote the empirical minimizer over the class $\mathcal{F}$. Argue that it is likely to have good performance, that is, for all $\epsilon \ge 0$ we have

$$\mathbb{P}\left(R(\widehat{f}_n) \ge R(f^\star) + 2\epsilon \right) \le \mathrm{card}(\mathcal{F}) \cdot \exp\left(-\frac{1}{2} \min\left\{ \frac{5}{6} \frac{n\epsilon^2}{R^\star(1 - R^\star) + 2\epsilon(1 - 2\epsilon)}, \frac{n\epsilon}{2} \right\} \right).$$

(d) Using the result of part (c), argue that with probability at least $1 - \delta$,

$$R(\widehat{f}_n) \le R(f^\star) + \frac{48 \log \frac{|\mathcal{F}|}{\delta}}{5n} + 4\sqrt{\frac{3}{5}} \cdot \frac{\sqrt{R^\star(1 - R^\star) \cdot \log \frac{|\mathcal{F}|}{\delta}}}{\sqrt{n}}.$$

Why is this better than an inequality based purely on the boundedness of the loss ℓ, such as Theorem 4.4 or Corollary 4.6? What happens when there is a perfect risk minimizer $f^\star$?

Question 2.5 (Mixtures are as good as point distributions)**:** Let P be a Laplace(λ) distribution on $\mathbb{R}$, meaning that $X \sim P$ has density

$$p(x) = \frac{\lambda}{2} \exp(-\lambda |x|).$$

Assume that $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} P$, and let P^n denote the n-fold product of P. In this problem, we compare the predictive performance of distributions from the normal location family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with the mixture distribution Q^π over $\mathcal{P}$ defined by the normal prior distribution $\mathsf{N}(\mu, \tau^2)$, that is, $\pi(\theta) = (2\pi\tau^2)^{-1/2} \exp(-(\theta - \mu)^2/2\tau^2)$.

(a) Let $P_{\theta,\Sigma}$ be the multivariate normal distribution with mean $\theta \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$. What is $D_{\mathrm{kl}}\left(P^n \| P_{\theta,\Sigma}\right)$?

(b) Show that $\inf_{\theta \in \mathbb{R}^n} D_{\mathrm{kl}}\left(P^n \| P_{\theta,\Sigma}\right) = D_{\mathrm{kl}}\left(P^n \| P_{0,\Sigma}\right)$, that is, the mean-zero normal distribution has the smallest KL-divergence from the Laplace distribution.

(c) Let Q_n^π be the mixture of the n-fold products in $\mathcal{P}$, that is, Q_n^π has density

$$q_n^\pi(x_1^n) = \int_{-\infty}^{\infty} \pi(\theta) p_\theta(x_1) \cdots p_\theta(x_n) d\theta,$$

where π is $\mathsf{N}(0, \tau^2)$. What is $D_{\mathrm{kl}}\left(P^n \| Q_n^\pi\right)$?

(d) Show that the redundancy of Q_n^π under the distribution P is asymptotically nearly as good as the redundancy of any $P_\theta \in \mathcal{P}$, the normal location family (so P_θ has density $p_\theta(x) = (2\pi\sigma^2)^{-1/2} \exp(-(x - \theta)^2/2\sigma^2)$). That is, show that

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_P \left[\log \frac{1}{q_n^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = \mathcal{O}(\log n)$$

for any prior variance $\tau^2 > 0$ and any prior mean $\mu \in \mathbb{R}$, where the big-Oh hides terms dependent on τ^2, σ^2, μ^2.

(e) **Extra credit:** Can you give an interesting condition under which such redundancy guarantees hold more generally? That is, using Proposition 9.7 in the notes, give a general condition under which

$$\mathbb{E}_P \left[\log \frac{1}{q^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = o(n)$$

as $n \to \infty$, for all $\theta \in \Theta$.