Stats 231 / CS229T Potential Problems

Contents

1 Background and (sort of) basics

1.1 Probability

Question 1: Let $X_i \in \mathbb{R}$ be i.i.d. according to a distribution with CDF F, which for simplicity we assume to be continuous. Let F_n be the empirical CDF given by $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le t\}$. Without appealing to the Glivenko-Cantelli theorem, show that

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \overset{p}{\to} 0.$$

Hint: Use the fact that F and F_n are non-decreasing and consider subsets of $\mathbb{R}$.

Question 2 (Moment generating function background)**:** A mean zero random variable X is σ^2-*sub-Gaussian* if $\mathbb{E}[\exp(\lambda X)] \le \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbb{R}$.

(a) Show that if Z is mean-zero Gaussian with variance σ^2, then $\mathbb{E}[\exp(\lambda Z)] = \exp(\frac{\lambda^2\sigma^2}{2})$.

(b) Show that if X_i, $i = 1, \ldots, n$, are i.i.d. mean zero σ^2-sub-Gaussian random variables, then $\mathbb{E}[\max_{i \le n} X_i] \le \sqrt{2\sigma^2 \log n}$.

Question 3 (Moment generating functions of squares)**:** In this question, we investigate sub-exponential and sub-Gaussian random variables. We let $[t]_+ = \max\{0, t\}$ denote the positive part, and say that $1/0 = +\infty$.

(a) Let Z be $\mathsf{N}(0, \sigma^2)$. Show that

$$\mathbb{E}[e^{\lambda Z^2}] = \frac{1}{\sqrt{[1 - 2\lambda\sigma^2]_+}}.$$

(b) Let X be a mean-zero σ^2-sub-Gaussian random variable. Show that

$$\mathbb{E}[e^{\lambda X^2}] \le \frac{1}{\sqrt{[1 - 2\lambda\sigma^2]_+}} \quad \text{for } \lambda \ge 0.$$

Hint: Introduce an independent Gaussian Z (with some particular variance) and compute $\mathbb{E}[e^{ZX}]$.

(c) Let $Z \sim \mathsf{N}(0, \sigma^2)$. Show that $Z^2 - \mathbb{E}[Z^2]$ is sub-exponential and give sub-exponential parameters for it.

Question 4 (An independent error bound)**:** You have a classifier that has a probability α of making a mistake on a random example drawn from some probability distribution P. You run the classifier on n i.i.d. examples from P. You want to ensure that the classifier errs at least once with probability at least $1 - \delta$. How large does n have to be (as a function of α and δ) for this to happen?

Question 5 (Asymptotics)**:** Suppose we have two sequences of i.i.d. random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. All $2n$ random variables are jointly independent and each random variable has mean μ and variance σ^2. Define the average difference:

$$D_n = \frac{1}{n}\sum_{i=1}^n (X_i - Y_i).$$

Compute the probability that D_n deviates by some amount determined by some $c > 0$:

(a) $\lim_{n\to\infty} \mathbb{P}[D_n \ge c]$

(b) $\lim_{n\to\infty} \mathbb{P}[D_n \ge \frac{c}{\sqrt{n}}]$.

Express your answers in terms of the cumulative density function (CDF) Φ of the standard normal distribution ($\Phi(z) = \mathbb{P}[Z \le z]$ for $Z \sim \mathsf{N}(0, 1)$).

Question 6 (Moments)**:**

(a) The *variance* of a random variable X is $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Show that

$$\mathrm{Var}(X) = \inf_{b\in\mathbb{R}} \mathbb{E}[(X - b)^2] = \mathbb{E}[(X - \mathbb{E}[X])^2],$$

that is, $\mathbb{E}[X]$ minimizes $\mathbb{E}[(X - b)^2]$ over all $b \in \mathbb{R}$.

(b) Suppose X_1 and X_2 are two independent random variables with variances σ_1^2 and σ_2^2. Compute $\mathrm{Var}(\alpha X_1 + \beta X_2)$ for $\alpha, \beta \in \mathbb{R}$.

(c) Let X and Y be real-valued random variables and f be some arbitrary function function. Show that the following decomposition holds:

$$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y \mid X])^2] + \mathbb{E}[\mathrm{Var}(Y \mid X)]$$

where

$$\mathrm{Var}(Y \mid X) = \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X]$$

is the variance of Y conditioned on X.

1.2 Convex analysis

Question 7 (Convexity)**:** A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ we have $f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda)f(y)$. You may use that if f is twice continuously differentiable, then f is convex if and only if $\nabla^2 f(x) \succeq 0$, that is, the Hessian $\nabla^2 f$ is positive semi-definite.

(a) Show that $f(x) = a^T x + b$ is convex for any $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

(b) Show that if $A \in \mathbb{R}^{n\times m}$ is a matrix and $b \in \mathbb{R}^n$ is a vector, then $f(Ax + b)$ is convex whenever f is.

(c) Show that the function $f(t) = \log(1 + e^{-t})$ is convex.

(d) Show that if $f_1, f_2, \ldots, f_k$ are convex functions, then $f(x) = \max\{f_1(x), \ldots, f_k(x)\}$ is convex.

Question 8 (Subgradients)**:** The subgradient set of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at a point x is defined by

$$\partial f(x) := \left\{g \in \mathbb{R}^n \mid f(y) \ge f(x) + g^T(y - x) \text{ for all } y \in \mathbb{R}^n\right\}.$$

It is a theorem that at any point in the interior of its domain, a convex function f has a non-empty subgradient set, and moreover, $\partial f(x) = \{\nabla f(x)\}$ at all points where f is differentiable.

(a) Draw a picture of a convex function with at least one point where it is non-differentiable, and draw lines defined by some of the linear functions $\widehat{f}(y) = f(x) + g^T(y - x)$ for $g \in \partial f(x)$.

(b) Let $f(x) = |x|$. Give formulae for $\partial f(x)$ for all $x \in \mathbb{R}$.

(c) Let $f(x) = \max\{0, x\}$. Give formulae for $\partial f(x)$ for all $x \in \mathbb{R}$.

(d) Let $f(x) = \frac{1}{2}x^2$. Give formulae for $\partial f(x)$ for all $x \in \mathbb{R}$.

(e) Let $f(x) = g(Ax + b)$, where g is convex. Show that

$$\partial f(x) \subset A^T \partial g(Ax + b),$$

where $A\mathcal{X} = \{Ax : x \in \mathcal{X}\}$ for a set $\mathcal{X}$. (Generally, this containment is an equality.)

Question 9 (Subgradients and convexity)**:** Consider the prediction problem of mapping some input $x \in \mathbb{R}^d$ to output y (in regression, we have $y \in \mathbb{R}$; in classification, we have $y \in \{-1, +1\}$). A linear predictor is governed by a weight vector $w \in \mathbb{R}^d$, and we typically wish to choose w to minimize the cumulative loss over a set of training examples. Two popular loss functions for classification and regression are defined (on a single example (x, y)) as follows:

- Squared loss: $\ell(w; x, y) = \frac{1}{2}(y - w \cdot x)^2$.
- Hinge loss: $\ell(w; x, y) = \max\{1 - yw \cdot x, 0\}$.

Let's study some properties of these loss functions. These will be used throughout the entire class, so it's important to obtain a good intuition for them.

(a) Show that each of the two loss functions is convex. *Hint:* whenever possible, use the compositional properties of convexity (i.e., sum of two convex functions is convex, etc.).

(b) Compute the subgradient of each of the two loss functions with respect to w.

(c) Suppose that $|y| \le 1$, $\|w\|_2 \le B$, and $\|x\|_2 \le C$ for some constants $B, C < \infty$. Give bounds on the ℓ_2-norms $\|\cdot\|_2$ of the subgradients g of each of the losses.

(In this class, many of the generalization bounds rely on control of the norms of the gradients, so it's important to get a feel for these dependencies.)

Question 10 (Exponential families)**:** Recall that an exponential family is a collection of probability distributions over $x \in \mathcal{X}$ (for simplicity, assume $\mathcal{X}$ is finite), parameterized by $\theta \in \mathbb{R}^d$ and a *sufficient statistic* (feature vector) $\phi : \mathcal{X} \to \mathbb{R}^d$. The density (probability mass function) of the exponential family has the form

$$p(x; \theta) = \exp\left(\theta^T \phi(x) - A(\theta)\right),$$

and $A(\theta) = \log \sum_{x \in \mathcal{X}} \exp(\theta^T \phi(x))$ is the log-partition function.

(a) Compute $\nabla A(\theta)$.

(b) Compute $\nabla^2 A(\theta)$.

(c) Give a probabilistic interpretation of each of these quantities.

(d) Argue that $A(\theta)$ is convex in θ.

1.3 Linear algebra

Question 11 (Linear algebra)**:**

(a) In linear regression, we are given a design matrix $X \in \mathbb{R}^{n\times d}$ where each row corresponds to a data point, and a vector of responses $Y \in \mathbb{R}^n$. Define the estimator as follows:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|X\theta - Y\|_2^2 + \lambda \|\theta\|_2^2 .$$

Assume $\lambda > 0$. Compute the closed form solution for $\widehat{\theta}$.

The *dual norm* $\|\cdot\|_*$ of a norm $\|\cdot\|$ is

$$\|w\|_* := \sup_{v:\|v\|\le 1} v^T w.$$

(b) The ℓ_1-norm $\|\cdot\|_1$ of a vector $v \in \mathbb{R}^n$ is $\|v\|_1 = \sum_{j=1}^n |v_j|$. Compute the dual norm of the ℓ_1-norm.

(c) The Cauchy-Schwartz inequality is that

$$u^T v \le \|u\|_2 \|v\|_2$$

for any u, v. Using that $\|\alpha u + \beta v\|_2^2 \ge 0$ for all $\alpha, \beta \in \mathbb{R}$, prove the Cauchy-Schwartz inequality.

(d) Compute the dual norm of the ℓ_2-norm.

(e) Show for any $x, y \in \mathbb{R}$ and any $p \in [1,\infty]$ with $q \in [1,\infty]$ such that $1/p + 1/q = 1$ that

$$xy \le \frac{\eta^p}{p}|x|^p + \frac{1}{\eta^q q}|y|^q$$

for all $\eta \ge 0$. [*Hint:* Either use the concavity of the logarithm or minimize the preceding expression in η]

(f) For $p \in [1,\infty]$, compute the dual norm of the ℓ_p-norm where $\|v\|_p = (\sum_{j=1}^n |v_j|^p)^{1/p}$. [*Hint:* Give an upper bound on $\sum_{j=1}^n v_j u_j$ and minimize it.]

(g) The *nuclear norm* of a matrix $A \in \mathbb{R}^{m\times n}$ is $\|A\|_* := \sum_{i=1}^{m\wedge n} |\sigma_i(A)|$, where the $\sigma_i(A)$ are the singular values of A. Show that the nuclear norm of a symmetric positive semi-definite matrix A is equal to its trace ($\mathrm{tr}(A) = \sum_{i=1}^n A_{ii}$). (For this reason, the nuclear norm is sometimes called the trace norm.) [*Hint*: use the fact that $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$.]

(h) **Hard but fun:** The ℓ_2-*operator norm* of a matrix A, $\|A\|_{\mathrm{op}}$, is its maximum singular value. Show that the nuclear and operator norms are dual to one another when we define the inner product between $m \times n$ matrices by $\langle A, B\rangle = \mathrm{tr}(A^T B)$.

2 Concentration and generalization

Question 12 (Concentration inequalities)**:** Let X_i be independent random variables with $|X_i| \le c$ and $\mathbb{E}[X_i] = 0$.

(a) Let $\sigma_i^2 = \mathrm{Var}(X_i)$. Prove that

$$\mathbb{E}[e^{\lambda X}] \le \exp\left(\frac{\sigma_i^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right).$$

(b) Let $h(u) = (1+u)\log(1+u) - u$ and let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Prove *Bennett's inequality*, that is, for any $t \ge 0$ we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right).$$

(c) Under the notation of part (b), prove *Bernstein's inequality*, that is, that for any $t \ge 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge t\right) \vee \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \le -t\right) \le \exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right),$$

where $a \vee b = \max\{a, b\}$.

(d) When is Bernstein's inequality tighter than the Hoeffding's inequality for bounded random variables? Recall that Hoeffding's inequality states (under the above conditions on X_i) that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge t\right) \le \exp\left(-\frac{nt^2}{2c^2}\right).$$

Question 13: In the realizable setting with binary classification (where the expected risk minimizer $h^\star$ satisfies $L(h^\star) = 0$ for the 0-1 error), we obtained excess risk bounds of $O(1/n)$, but in the unrealizable setting, we had $O(\sqrt{1/n})$. What if the learning problem is *almost* realizable, in that $L(h^\star)$ is small? This problem explores ways to interpolate between $1/n$ and $1/\sqrt{n}$ rates, showing that (roughly) $\sqrt{L(h^\star)/n} + 1/n$ rates are possible by developing generalization bounds that depend on the *variance* of losses (recall Question 12).

(a) Assume that the loss function $\ell(y, t)$ takes values in $[0, 1]$, where $L(h) = \mathbb{E}[\ell(Y, h(X))]$, and let $\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, h(X_i))$. Show that for all $\epsilon \ge 0$ we have

$$\mathbb{P}\left(\widehat{L}_n(h) - L(h) \ge \epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2(L(h) + \epsilon/3)}\right).$$

(Note that if $L(h) = 0$, this bound scales as $e^{-n\epsilon} \ll e^{-n\epsilon^2}$ for $\epsilon \approx 0$.)

(b) We now show that bad hypotheses usually look pretty bad. Fix any $\varepsilon(h), \epsilon \ge 0$, and assume that

$$L(h) \ge \varepsilon(h) + \epsilon.$$

Show that

$$\mathbb{P}\left(\widehat{L}_n(h) \le \varepsilon(h)\right) \le \exp\left(-\frac{n\epsilon^2}{2(\varepsilon(h) + 4\epsilon/3)}\right).$$

(c) Assume $\text{card}(\mathcal{H}) < \infty$ and let $h^\star$ satisfy $L(h^\star) = \min_{h \in \mathcal{H}} L(h)$. Using the preceding parts, conclude that if $\widehat{h}_n \in \text{argmin}_{h \in \mathcal{H}} \widehat{L}_n(h)$, then

$$\mathbb{P}\left(L(\widehat{h}_n) - L(h^\star) \geq 2\epsilon\right) \leq \text{card}(\mathcal{H}) \exp\left(-\frac{n\epsilon^2}{2(L(h^\star) + 7\epsilon/3)}\right).$$

Show that this implies (for appropriate numerical constants c_1, c_2) that with probability at least $1 - \delta$, we have

$$L(\widehat{h}_n) \leq L(h^\star) + c_1 \sqrt{\frac{L(h^\star) \log \frac{\text{card}(\mathcal{H})}{\delta}}{n}} + c_2 \frac{\log \frac{\text{card}(\mathcal{H})}{\delta}}{n}.$$

(d) How does this bound compare with a more naive strategy based on applying Hoeffding's inequality and a union bound?

Question 14 (VC Dimension)**:**

(a) Let $\mathcal{X} = \mathbb{R}^2$ and consider the hypothesis class of indicators for convex polygons, that is,

$$\mathcal{H} = \{h_C(x) = \mathbf{1}\{x \in C\} : C \text{ is a convex polygon}\}.$$

What is $\mathsf{VC}(\mathcal{H})$?

(b) A decision tree T is a binary tree that classifies points in $\mathbb{R}^d$. Each internal node (non-leaf node) v in T has an attribute $j_v \in \{1, 2, \ldots, d\}$ and a threshold $t_v \in \mathbb{R}$. Each leaf node is labeled with one of the two classes, +1 or -1. Given a point $x \in \mathbb{R}^d$, we start from the root, and every time we encounter an internal node v, we check the condition $\mathbf{1}\{x_{j_v} \geq t_v\}$. We go to the left child if the condition is not met, and the right child otherwise. We repeat such process until we reach a leaf node, and classifies the point according to the label of the node.

Show that the VC dimension of the hypothesis class corresponding to all depth-k decision trees defined above is $\Omega(2^k \log d)$.

Question 15 (Rademacher complexity)**:** In many applications, for example, in natural language processing (NLP), one has very sparse feature vectors in very high dimensions. Suppose that we know that any feature vector $x \in \{0, 1\}^d$ satisfies $\|x\|_1 \leq k$, i.e. there are at most k non-zeros.

(a) Give an example application and data representation where such characteristics might hold.

You decide to use a linear classifier for this "sparse x" problem, where you represent the classifier by a weight vector $w \in \mathbb{R}^d$ so that $f(x) = w^\top x$, and you restrict your classifiers to be in a particular norm ball $\{w : \|w\| \leq B\}$.

(b) Is using the ℓ_1-norm ball, i.e. $\mathcal{F} = \{x \mapsto f(x) = w^\top x : \|w\|_1 \leq B\}$ likely to be a good idea? In a sentence or two, explain why or why not. (No need for serious mathematical derivations.)

(c) You decide instead to use dense feature vectors, restricting w to an ℓ_∞ norm ball, i.e.

$$\mathcal{F} := \{f \mid f(x) = w^\top x, \|w\|_\infty \leq B\}.$$

Give an upper bound on $R_n(\mathcal{F})$, which should depend on k (the number of non-zeros), n, B, and d.

Question 16 (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 $-f \in \mathcal{F}$,

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

Question 17 (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

$$\mathbb{E}[X_t] = \mathbb{E}[Y_t] = 0 \;\text{ and }\; \mathbb{E}[(X_t - X_s)^2] \le \mathbb{E}[(Y_t - Y_s)^2] \;\text{ for all } s, t \in T \tag{1}$$

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 (1) 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 (1) 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 $g_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$ be independent standard Gaussians and $Z_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ 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)

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

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

(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?

Question 18 (Multiclass Gaussian complexity)**:** In multiclass classification problems (i.e. there are $k \ge 3$ classes), a natural margin-based formulation—analogous to the formulation for binary problems—is to have $\phi : \mathbb{R}^k \to \mathbb{R}$ be convex, where ϕ is symmetric and increasing in its last $k-1$ arguments and non-increasing in its first argument. As examples, we might take

$$\phi_{\log}(v) = \log\left(\sum_{l=1}^k e^{v_l - v_1}\right) \quad \text{or} \quad \phi_{\text{hinge}}(v) = [1 - v_1]_+ + \sum_{l=2}^k [1 + v_l]_+ .$$

We would like to learn a weight vector $w_l \in \mathbb{R}^d$ for each class $l \in \{1, \ldots, k\}$, so that given a label $y \in \{1, \ldots, k\}$ and point x we classify the pair (x, y) correctly if $w_y^T x > w_l^T x$ for all $l \ne y$. Let e_l denote the lth standard basis vector. Given a label $y \in \{1, \ldots, k\}$, define the permutation matrix

$$\Pi_y = [e_y \; e_1 \; e_2 \; \cdots \; e_{y-1} \; e_{y+1} \; \cdots \; e_k] \in \{0,1\}^{k \times k} \quad \text{so} \quad \Pi_y v = \begin{bmatrix} v_y & v_1 & \cdots & v_{y-1} & v_{y+1} & \cdots & v_k \end{bmatrix}^T,$$

that is, Π_y moves the yth position to the first coordinate and shifts the others approriately. We then define the loss of the matrix $W = [w_1 \; \cdots \; w_k] \in \mathbb{R}^{d \times k}$ on the pair (x, y) by $\ell(W; x, y) = \phi(\Pi_y W^T x)$.

(a) In about one sentence, explain why this choice of loss may be a good idea.

(b) Show that each of $\phi_{\log}$ and ϕ_{hinge} are convex.

(c) Give explicit formulae for $\ell(W; x, y) = \phi_{\log}(\Pi_y W^T x)$ and $\phi_{\text{hinge}}(\Pi_y W^T x)$.

(d) Let $\mathcal{F} = \{\ell(W; \cdot) - \ell(W^\star; \cdot) \mid W \in \mathcal{W}\}$ be a centered loss class for a loss of the form $\ell(W; x, y) = \phi(\Pi_y W^T x)$, where $\mathcal{W} \subset \mathbb{R}^{d \times k}$. Assume also that ϕ is M-Lipschitz with respect to the ℓ_2-norm. Show that the empirical Gaussian complexity of $\mathcal{F}$ satisfies

$$\widehat{G}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{W \in \mathcal{W}} \sum_{i=1}^n g_i \left(\ell(W; x_i, y_i) - \ell(W^\star; x_i, y_i)\right)\right] \le M\mathbb{E}\left[\sup_{W \in \mathcal{W}} \sum_{i=1}^n Z_i^T (W - W^\star)^T x_i\right]$$

for $Z_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_k)$ and $g_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$, where $W^\star \in \mathbb{R}^{d \times k}$ is some fixed matrix.

(e) Suppose that $\mathcal{W} = \{W \in \mathbb{R}^{d \times k} \mid \|W - W^\star\|_{\rm Fr} \le r\}$, where $\|\cdot\|_{\rm Fr}$ denotes the Frobenius norm. Give an upper bound on $\widehat{G}_n(\mathcal{F})$ from part (d) depending only on a numerical constant and (possibly a subset of) the terms n, d, k, M, r.

(f) Suppose that

$$\mathcal{W} = \{W = [w_1 \ \cdots \ w_k] \in \mathbb{R}^{d \times k} \mid \|w_l\|_1 \le r \text{ for } l = 1, \ldots, k\}$$

and let $W^\star = 0$. Give an upper bound on $\widehat{G}_n(\mathcal{F})$ from part (d) depending only on a numerical constant and (possibly a subset of) the terms n, d, k, M, r.

3 Online learning and stochastic optimization

Question 19 (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

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

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 $D_h(x, y) \ge \frac{1}{2} \|x - y\|^2$ for all $x, y \in X$.

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

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

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

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

(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

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

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

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

Question 20 (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 $G_t = \sum_{\tau=1}^{t} \mathrm{diag}(g_\tau)^2$ 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 $x^\star \in X$,

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

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

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

(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?

Question 21 (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

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

Question 22 (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

$$\inf_{x \in X} \sup_{y \in Y} x^T A y = \sup_{y \in Y} \inf_{x \in X} x^T A y. \tag{2}$$

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 $\|A\|_{\rm op} \le 1$ and $\|x - x'\|_2 \le 1$, $\|y - y'\|_2 \le 1$ 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

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

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 (2). (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 $T \to \infty$.)

4 Kernels and representations

Question 23: Let $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a valid kernel function. Define

$$\mathsf{k}_{\text{norm}}(x, z) := \frac{\mathsf{k}(x, z)}{\sqrt{\mathsf{k}(x, x)}\sqrt{\mathsf{k}(z, z)}}.$$

Is k_{norm} a valid kernel? Justify your answer.

Question 24: Consider the class of functions

$$\mathcal{H} := \left\{f : f(0) = 0, f' \in L^2([0, 1])\right\},$$

that is, functions $f : [0, 1] \to \mathbb{R}$ with $f(0) = 0$ that are almost everywhere differentiable, where $\int_0^1 (f'(x))^2 dx < \infty$. On this space of functions, we define the inner product by

$$\langle f, g\rangle = \int_0^1 f'(x)g'(x)dx.$$

Show that $\mathsf{k}(x, z) = \min\{x, z\}$ is the reproducing kernel for $\mathcal{H}$, so that it is (i) positive semidefinite and (ii) a valid kernel.

Question 25: Consider the Sobolev space $\mathcal{F}_k$, which is defined as the set of functions that are $(k - 1)$-times differentiable and have kth derivative almost everywhere on $[0, 1]$, where the kth derivative is square-integrable. That is, we define

$$\mathcal{F}_k := \left\{f : [0, 1] \mid f^{(k)}(x) \in L^2([0, 1])\right\}.$$

We define the inner product on $\mathcal{F}_k$ by

$$\langle f, g\rangle = \sum_{i=0}^{k-1} f^{(i)}(x)g^{(i)}(x) + \int_0^1 f^{(k)}(x)g^{(k)}(x)dx.$$

(a) Find the representer of evaluation for this Hilbert space, that is, find a function $r_x : [0, 1] \to \mathbb{R}$ (defined for each $x \in [0, 1]$) such that $r_x \in \mathcal{F}_k$ and

$$\langle r_x, f\rangle = f(x)$$

for all x.

(b) What is the reproducing kernel $\mathsf{k}(x, z)$ associated with this space? (Recall that $\mathsf{k}(x, z) = \langle r_x, r_z\rangle$ for an RKHS.)

(c) Show that $\mathcal{F}_k$ is a Hilbert space, meaning that $\|f\|^2 = \langle f, f\rangle$ defines a norm and that $\mathcal{F}_k$ is complete for the norm.

Question 26: The variation distance between probability distributions P and Q on a space $\mathcal{X}$ is defined by $\|P - Q\|_{\text{TV}} = \sup_{A \subset \mathcal{X}} |P(A) - Q(A)|$.

(a) Show that

$$2\|P - Q\|_{\text{TV}} = \sup_{f : \|f\|_\infty \le 1} \{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\}$$

where the supremum is taken over all functions with $f(x) \in [-1, 1]$, and the first expectation is taken with respect to P and the second with respect to Q. You may assume that P and Q have densities.

Question 27: In a number of experimental situations, it is valuable to determine if two distributions P and Q are the same or different. For example, P may be the distribution of widgets produced by one machine, Q the distributions of widgets by a second machine, and we wish to test if the two distributions are the same (to within allowable tolerances). Let $\mathcal{H}$ be an RKHS of functions with domain $\mathcal{X}$ and reproducing kernel k, and let P and Q be distributions on $\mathcal{X}$.

(a) Let $\|\cdot\|_{\mathcal{H}}$ denote the norm on the Hilbert space $\mathcal{H}$. Show that

$$D_{\mathsf{k}}(P,Q)^2 := \sup_{f:\|f\|_{\mathcal{H}}\le 1} \left\{ |\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Z)]|^2 \right\} = \mathbb{E}[\mathsf{k}(X,X')] + \mathbb{E}[\mathsf{k}(Z,Z')] - 2\mathbb{E}[\mathsf{k}(X,Z)]$$

where $X, X' \stackrel{\text{iid}}{\sim} P$ and $Z, Z' \stackrel{\text{iid}}{\sim} Q$.

(b) A kernel $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *universal* if the induced RKHS $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ can arbitrarily approximate continuous functions. That is, for any $\phi : \mathcal{X} \to \mathbb{R}$ continuous and $\epsilon > 0$, there is some $f \in \mathcal{H}$ such that

$$\sup_{x\in\mathcal{X}} |f(x) - \phi(x)| \le \epsilon.$$

Show that if k is universal, then

$$D_{\mathsf{k}}(P,Q) = 0 \quad \text{if and only if} \quad P = Q.$$

You may assume $\mathcal{X}$ is a metric space and that $P = Q$ iff $P(A) = Q(A)$ for all compact $A \subset \mathcal{X}$.

(c) You wish to estimate $D_{\mathsf{k}}(P,Q)$ given samples from each of the distributions. Assume that $\mathsf{k}(x,z) \in [-B, B]$ for all $x, z \in \mathcal{X}$. Let $X_i \stackrel{\text{iid}}{\sim} P$, $i = 1, \ldots, n_1$ and $Z_i \stackrel{\text{iid}}{\sim} Q$, $i = 1, \ldots, n_2$. Define

$$\widehat{K}(X_{1:n_1}) := \binom{n_1}{2}^{-1} \sum_{1\le i<j\le n_1} \mathsf{k}(X_i, X_j), \quad \widehat{K}(Z_{1:n_2}) := \binom{n_2}{2}^{-1} \sum_{1\le i<j\le n_2} \mathsf{k}(Z_i, Z_j),$$

and

$$\widehat{K}(X_{1:n_1}, Z_{1:n_2}) := \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathsf{k}(X_i, Z_j).$$

Show that $\mathbb{E}[\widehat{K}(X_{1:n})] = \mathbb{E}[\mathsf{k}(X,X')]$ and $\mathbb{E}[\widehat{K}(X_{1:n_1}, Z_{1:n_2})] = \mathbb{E}[\mathsf{k}(X,Z)]$ for $X, X' \stackrel{\text{iid}}{\sim} P$ and $Z, Z' \stackrel{\text{iid}}{\sim} Q$. Show for some numerical constant $c > 0$ that for all $t \ge 0$,

$$\mathbb{P}\left(\left| \widehat{K}(X_{1:n}) - \mathbb{E}[\mathsf{k}(X,X')] \right| \ge t \right) \le 2\exp\left(-c\frac{nt^2}{B^2} \right)$$

and

$$\mathbb{P}\left(\left| \widehat{K}(X_{1:n_1}, Z_{1:n_2}) - \mathbb{E}[\mathsf{k}(X,Z)] \right| \ge t \right) \le 2\exp\left(-c\frac{n_1 t^2}{B^2} \right) + 2\exp\left(-c\frac{n_2 t^2}{B^2} \right).$$

(d) Define the empirical Hilbert distances

$$\widehat{D}^2_{\mathsf{k}}(P,Q) := \binom{n_1}{2}^{-1} \sum_{1\le i<j\le n_1} \mathsf{k}(X_i, X_j) + \binom{n_2}{2}^{-1} \sum_{1\le i<j\le n_2} \mathsf{k}(Z_i, Z_j) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathsf{k}(X_i, Z_j).$$

Show that for all $t \ge 0$,

$$\mathbb{P}\left(\left| \widehat{D}^2_{\mathsf{k}}(P,Q) - D^2_{\mathsf{k}}(P,Q) \right| \ge t \right) \le C\exp\left(-c\frac{\min\{n_1, n_2\} t^2}{B^2} \right)$$

where $0 < c$, $C < \infty$ are numerical constants.