Exam (2014) solutions

CS229T/STATS231 (Fall 2018–2019)

Note: please do not copy or distribute.

SUNet ID: _____

Name: _____

Collaborators: none

1. This exam is open notes and open laptop, but you must turn off the wireless.
2. Please write clearly and give rigorous justification for each non-trivial step.
3. You may cite results from the course notes without proof.
4. When we ask for an upper bound, to get full credit, in addition to being mathematically correct, you must have the right dependencies on all the problem-dependent quantities (e.g., if it is possible to get d, then $d \log d$ is not acceptable, but $2d$ is acceptable).

1. Ensembling linear predictors (*15 points*)

Consider binary classification with the hinge loss: $\ell((x,y),w) = \max\{0, 1-(w\cdot x)y\}$, where $x \in \mathbb{R}^d$ and $y \in \{-1,+1\}$.

Suppose we are given K positive definite $d \times d$ matrices $A_1, \dots, A_K$. For each $k = 1, \dots, K$, define a subset of the weight vectors:

$$S_k = \{w \in \mathbb{R}^d : w^\top A_k w \le 1\}.$$

Assume the inputs x are bounded according to $x^\top A_k^{-1} x \le 1$ for each $k = 1, \dots, K$.

Each S_k captures different prior knowledge about the weights w, but a priori, we probably don't know which one to use. In this problem, we'll try to combine the strengths of all K hypothesis classes.

a. (*5 points*)

Consider the online learning setting, where on each iteration t, the learner plays a weight vector w_t, and nature plays the loss function $f_t(w) = \ell((x_t, y_t), w)$.

For this part, fix a hypothesis class $k \in \{1, \dots, K\}$. Define the regret with respect to the k-th hypothesis class to be

$$\text{Regret}_k \stackrel{\text{def}}{=} \max_{u \in S_k} \sum_{t=1}^{T} [f_t(w_t) - f_t(u)].$$

Your task: define a regularizer $\psi(w)$ such that running online mirror descent with respect to ψ yields the following regret bound:

$$\text{Regret}_k \le \sqrt{T}.$$

You can use the fact that for positive definite A, $\|w\|_A \stackrel{\text{def}}{=} \sqrt{w^\top A w}$ is a norm with dual norm $\|w\|_{A^{-1}}$.

Solution:

Define the regularizer $\psi(w) = \frac{1}{2\eta}\|w\|_{A_k}^2$, which is clearly $\frac{1}{\eta}$-strongly convex with respect to the $\|\cdot\|_{A_k}$ norm. The dual norm is $\|\cdot\|_{A_k^{-1}}$. The main result about mirror descent states that

$$\text{Regret}_k(u) \le \frac{1}{2\eta}[u^\top A_k u - 0^\top A_k 0] + \frac{\eta}{2}\sum_{t=1}^{T} z_t^\top A_k^{-1} z_t.$$

For the hinge loss the subgradient is $z_t = a_t x_t$ for some $a_t \in \{-1, 0, +1\}$, so we have $z_t^\top A_k^{-1} z_t \le 1$. For $u \in S_k$, $u^\top A_k u \le 1$, so $\text{Regret}_k \le \frac{1}{2\eta} + \frac{\eta T}{2}$, which is upper bounded by $\sqrt{T}$ by setting $\eta = \frac{1}{\sqrt{T}}$.

b. (*5 points*)

Suppose we run K instances of the above online learning algorithms, each producing its own sequence of weights $w_1^k, w_2^k, \dots, w_T^k$ for $k = 1, \dots, K$.

Construct an algorithm that maintains a distribution $\alpha_t \in \Delta_K$ over the K instances. On each iteration t, your algorithm should (i) *deterministically* choose a weight vector $v_t \in \mathbb{R}^d$ and (ii) update the distribution α_t to α_{t+1} based on the vector of hinge losses $[f_t(w_t^1), \dots, f_t(w_t^K)]$ that you receive from nature. Using the result from part (a), show that your algorithm achieves the following regret bound:

$$\text{Regret} \stackrel{\text{def}}{=} \max_{u \in \cup_{k=1}^K S_k} \sum_{t=1}^{T} [f_t(v_t) - f_t(u)] \le (1 + \sqrt{2\log K})\sqrt{T}.$$

Note that your algorithm is now competing with the best weights over the union $\cup_{k=1}^K S_k$.

Solution:
Treat each of the K instances as an expert and apply the exponentiated gradient algorithm in the learning with expert advice setting. Let $z_t = [f_t(w_t^1), \ldots, f_t(w_t^K)] \in \mathbb{R}^K$ be the loss vector. Our algorithm will (i) choose the average weight vector

$$v_t = \sum_{k=1}^{K} \alpha_{t,k} w_t^k,$$

and (ii) update the distribution

$$\alpha_{t+1,k} \propto \alpha_{t,k} e^{-\eta z_{t,k}},$$

where $\eta = \sqrt{\frac{2\log(K)}{T}}$.

Now let us analyze the algorithm. Recall that if we randomly chose an instance k each iteration according to α_t, then we would suffer *expected loss* $\alpha_t \cdot z_t$. However, since the hinge loss f_t is convex, we have that $f_t(v_t) \leq \alpha_t \cdot z_t$. From the learning with expert advice result (observe that the hinge loss is non-negative), we get that

$$\max_{k=1}^{K} \sum_{t=1}^{T} [f_t(v_t) - f_t(w_t^k)] \leq \sqrt{2\log(K)T}.$$

By the bound from the part (a), we have that for each $k = 1, \ldots K$,

$$\max_{u \in S_k} \sum_{t=1}^{T} [f_t(w_t^k) - f_t(u)] \leq \sqrt{T}.$$

Adding the two inequalities yields the desired result.

c. (*5 points*)

Let us now leave the online setting and try ensembling in the batch setting. For each $k = 1, \ldots, K$, define the linear functions

$$\mathcal{F}_k \stackrel{\text{def}}{=} \{z \mapsto w \cdot z : w \in S_k\}.$$

Define the convex combinations:

$$\mathcal{F} \stackrel{\text{def}}{=} \left\{ \sum_{k=1}^{K} \alpha_k f_k : [\alpha_1, \ldots, \alpha_K] \in \Delta_K, \forall k\, f_k \in \mathcal{F}_k \right\}.$$

Show that the Rademacher complexity $R_n(\mathcal{F})$ can be upper bounded as follows:

$$R_n(\mathcal{F}) \leq \frac{K}{\sqrt{n}}.$$

Solution:

We have

$$
\begin{aligned}
R_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right] && (1)\\
&= \mathbb{E}\left[\sup_{k=1}^{K} \sup_{w_k\in S_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right] && (2)\\
&\leq \sum_{k=1}^{K} \mathbb{E}\left[\sup_{w_k\in S_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right] && (3)\\
&\leq \sum_{k=1}^{K} R_n(\mathcal{F}_k). && (4)
\end{aligned}
$$

To compute the Rademacher complexity of one $\mathcal{F}_k$, note that $w^\top A_k w \leq 1$ and $x^\top A_k^{-1} x \leq 1$ implies that $|w \cdot x| \leq 1$, so the same bound on the Rademacher complexity holds as in the L_2 constrained case: $R_n(\mathcal{F}_k) \leq \frac{1}{\sqrt{n}}$.

2. Intervals (*15 points*)

Suppose we are solving a binary classification problem with interval inputs $\mathcal{X} = \{[a, b] : a, b \in \mathbb{R}, a < b\}$ and output $\mathcal{Y} = \{-1, +1\}$.

a. (*5 points*)

Consider a hypothesis class indexed by sets S of at most k points:

$$\mathcal{H} = \{h_S : S \subset \mathbb{R}, |S| \leq k\},$$

where each hypothesis returns the number of points in S that intersects the interval x:

$$h_S(x) = |S \cap x|.$$

Show that the Rademacher complexity is upper bounded as:

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2k^3 \log(k(2n+1))}{n}}.$$

Solution:

Given n intervals, there are at most $2n$ left/right endpoints, and thus there are at most $2n+1$ regions between adjacent endpoints. For example, the two intervals $[1, 5]$ and $[2, 9]$ results in 5 regions: $(-\infty, 1], [1, 2], [2, 5], [5, 9], [9, \infty)$.

With a set of at most k points, at most k of these regions r can have $h_S(r) > 0$ (the rest have $h_S(r) = 0$); there are at most $(2n+1)^k$ ways to select these. Then, for each of these k regions r, $h_S(r) \in \{1, \ldots, k\}$, so there are k^k choices here. The values of the regions completely determine the values of the original intervals, so the number of possible vectors $|\{[h_S(x_1), \ldots, h_S(x_n)] : h \in \mathcal{H}\}|$ is upper bounded by $(2n+1)^k k^k$.

Applying Massart's finite lemma (note that $h_S([a, b])^2 \leq k^2 \stackrel{\text{def}}{=} M^2$), we get a bound on the empirical Rademacher complexity:

$$\hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2(k^2) \log(k(2n+1))^k}{n}}.$$

Taking expectations yields the result.

b. (*5 points*)

Suppose we obtained n i.i.d. training examples $\{(x_i, y_i)\}_{i=1}^n$, where each x_i is an interval. Define a loss function that looks at the fraction of points in S that fall into (out of) x_i when $y_i = -1$ ($y_i = +1$):

$$\ell((x_i, y_i), h) = \frac{1}{k} \begin{cases} k - h(x_i) & \text{if } y_i = +1 \\ h(x_i) & \text{if } y_i = -1. \end{cases}$$

Show that with probability at least $1 - \delta$, the expected risk of the empirical risk minimizer $\hat{h}$ does not exceed the empirical risk by too much:

$$L(\hat{h}) - \hat{L}(\hat{h}) \leq \frac{2R_n(\mathcal{H})}{k} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Solution:

First, note that

$$L(\hat{h}) - L(\hat{h}) \leq G_n \stackrel{\text{def}}{=} \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$$

. We have that $\mathbb{P}[G_n \geq \mathbb{E}[G_n] + \epsilon] \leq \exp(-2n\epsilon^2) \stackrel{\text{def}}{=} \delta$. Rewriting, we get that

$$L(\hat{h}) - \hat{L}(\hat{h}) \leq \mathbb{E}[G_n] + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

From McDiarmid's inequality and symmetrization, we have that $\mathbb{E}[G_n] \leq 2R_n(\mathcal{A})$, where $\mathcal{A}$ is the loss class. Conditioned on the dataset, we have a bound on the empirical Rademacher complexity of the loss class:

$$\begin{aligned}
\hat{R}_n(\mathcal{A}) &= \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\frac{1+y_i}{2} - \frac{h(x_i)y_i}{k}\right) \mid x_{1:n}, y_{1:n}\right] && (5)\\
&= \frac{1}{k}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i) y_i \mid x_{1:n}, y_{1:n}\right] && (6)\\
&= \frac{1}{k}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i) \mid x_{1:n}, y_{1:n}\right] && (7)\\
&= \frac{1}{k}\hat{R}_n(\mathcal{H}). && (8)
\end{aligned}$$

We used the fact that $y_i \in \{-1, +1\}$ is a constant, which can be folded into the random σ_i.

c. (*5 points*)

For an interval $x = [a, b]$, define its length to be $\text{len}(x) \stackrel{\text{def}}{=} b - a$. Show that the following is a valid kernel:

$$k(x, x') = \text{len}(x \cap x') + \text{len}(x)\text{len}(x').$$

Note that the intersection of an interval is an interval or the empty set (which has length 0).

Solution:
Define $k_1(x, x') = \text{len}(x \cap x')$ and $k_2(x, x') = \text{len}(x)\text{len}(x')$. Clearly, k_2 is a kernel (corresponding to the feature map $x \mapsto \text{len}(x)$) and a sum of kernels is a kernel, so it suffices to show that k_1 is a kernel.

Fix n intervals $x_1, \ldots, x_n$. Sort the left/right endpoints of all the intervals and let R be the regions between successive endpoints. Note that $|R| \leq 2n + 1$. Now, define a feature matrix $\Phi \in \mathbb{R}^{n\times|R|}$ to be

$$\Phi_{i,j} = \mathbb{I}[r_j \subset x_i]\sqrt{\text{len}(r_j)}.$$

It is easy to verify that the kernel matrix is $K = \Phi\Phi^\top$, which is positive semidefinite by construction.

3. Simple covariance matrices (*15 points*)

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be a n points which are elementwise bounded ($\|x_i\|_\infty \leq B$). Recall that the covariance matrix is defined as

$$\Sigma \stackrel{\text{def}}{=} \mathbb{E}\left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top\right].$$

a. (*5 points*) Suppose we obtain $x_1, \ldots, x_n$ drawn i.i.d. from some distribution p^*. Assume that $\mathbb{E}_{x \sim p^*}[x] = 0$. Define the empirical covariance

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top.$$

For each $\epsilon > 0$, derive an upper bound for

$$\mathbb{P}\left[\max_{1 \leq i,j \leq d} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \geq \epsilon\right].$$

Solution:
When x are zero-mean random vectors, one has

$$\mathbb{E}\left[\hat{\Sigma}\right] = \Sigma,$$

and hence

$$\hat{\Sigma} - \Sigma = \frac{1}{n} \sum_{i=1}^{n} [x_i x_i^\top - \mathbb{E}\left[x_i x_i^\top\right]].$$

That said, each entry of $\hat{\Sigma} - \Sigma$ is an average of n i.i.d. random variables each bounded in magnitude by $|x_{i,j} x_{i,j'}| \leq B^2$. By the two-sided Hoeffding's inequality, for each $1 \leq i, j \leq d$ we can bound

$$\mathbb{P}\left[\left|\hat{\Sigma}_{ij} - \Sigma_{ij}\right| \geq \epsilon\right] \quad \leq \quad 2\exp\left(-\frac{n\epsilon^2}{2B^4}\right).$$

Applying the union bound over all $1 \leq i, j \leq d$ gives

$$\begin{aligned}
\mathbb{P}\left[\max_{1 \leq i,j \leq d} \left|\hat{\Sigma}_{ij} - \Sigma_{ij}\right| \geq \epsilon\right] \quad &\leq \quad \sum_{1 \leq i,j \leq d} \mathbb{P}\left[\left|\hat{\Sigma}_{ij} - \Sigma_{ij}\right| \geq \epsilon\right] \\
&\leq \quad 2d^2 \exp\left(-\frac{n\epsilon^2}{2B^4}\right).
\end{aligned}$$

b. (*5 points*)

Now let's estimate the covariance matrix in an online fashion. Each iteration t, the learner predicts Σ_t, nature plays $x_t \in \mathbb{R}^d$, and the learner suffers loss $\frac{1}{2}\|\Sigma_t - x_t x_t^\top\|_F^2$, measured in the Frobenius norm. Construct an algorithm for predicting Σ_t and show that its regret can be upper bounded according to:

$$\text{Regret} \leq 4d^2 B^4 (\log T + 1).$$

Solution:

When the loss function takes the form

$$f_t(\Sigma) = \frac{1}{2}\left\|\Sigma - x_t x_t^\top\right\|_F^2.$$

FTL update has a closed-form solution

$$\Sigma_t = \frac{1}{t-1}\sum_{i=1}^{t-1} x_i x_i^\top.$$

Since

$$\left\|\mathrm{vec}(x_i x_i^\top)\right\|_2 \le \sqrt{d^2}B^2 = dB^2,$$

applying results in Page 17 of the course notes suggests that

$$\text{Regret} \le 4d^2B^4(\log T + 1).$$

c. (*5 points*) Let $x_1, \ldots, x_n$ be n points, and define the kernel $k(x, x') = x \cdot x'$. In practice, the data points do not necessarily have zero mean, and hence let us estimate the covariance matrix via the sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Show that $\mathrm{tr}(\hat{\Sigma})$ can be computed in a way that only depends on $x_1, \ldots, x_n$ through the kernel matrix.

Solution:
The claim follows from the following derivation:

$$\begin{aligned}
\mathrm{tr}\left(\hat{\Sigma}\right) &= \mathrm{tr}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\top\right\} \\
&= \mathrm{tr}\left\{\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - \overline{x}\overline{x}^\top\right\} && (9) \\
&= \frac{1}{n}\mathrm{tr}\left\{\sum_{i=1}^{n} x_i^\top x_i\right\} - \mathrm{tr}\left\{\overline{x}^\top \overline{x}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^\top x_i - \frac{1}{n^2}(x_1 + \cdots + x_n)^\top (x_1 + \cdots + x_n) && (10) \\
&= \frac{1}{n}\sum_{i=1}^{n} k(x_i, x_i) - \frac{1}{n^2}\sum_{1 \le i,j \le n} k(x_i, x_j). && (11)
\end{aligned}$$