Final Exam

SDS 391P.6, Spring 2026

Pratik Patil

Due: May 06 (Wednesday)

0 Guidelines

- The exam duration is 180 minutes. There are three main problems, worth 10 points each. Roughly speaking, we expect each main problem to be doable in about 30 minutes. This leaves about 30 minutes of buffer time per problem.
- Parts labeled [Bonus] are optional and are not required for full credit.
- The exam is closed book. You may use your midterm note sheet and one additional final-exam note sheet (front and back).
- We will provide loose white sheets in class. Please write your name and EID clearly on every page, and number your pages.
- Please begin each main problem on a new page. Show enough work that your reasoning can be followed.
- At the end of the exam, please staple your pages together. A stapler will be provided in the classroom. On the front page, please write the total number of pages in your submission.
- You may quote any result proved in lecture or homework, provided you state clearly what you are using. However, if a subpart explicitly asks you to prove or derive a result, then you should give the argument rather than only quote the result.
- Throughout the exam, constants denoted by $C, c, c_0, C_0, \ldots$ are positive absolute constants whose values may change from line to line.
- As with the midterm, we will allow post-exam completion. Any clearly labeled post-exam additions submitted by the announced deadline may earn up to 60% of the listed points on the relevant subpart. The recorded score on each subpart will be the maximum of the in-class score and the post-exam score. Post-exam additions must be your own work. They are meant as retrospective corrections of parts that you missed or did suboptimally.

1 Learning a Gaussian mixture model and covariance estimation

One of the simplest models of structured high-dimensional data is a *Gaussian mixture model*. In the two-cluster version, one observes points drawn from one of two Gaussian distributions with different means. A basic example is

$$X = G + \theta\, t\, u,$$

where $d \geqslant 2$, $u \in S^{d-1}$ is a fixed unit vector, $t > 0$ controls the separation between the two clusters, $G \sim \mathcal{N}(0, I_d)$, and $\theta \in \{-1, +1\}$ is a Rademacher random variable independent of G. Equivalently, X is drawn from $N(tu, I_d)$ or from $N(-tu, I_d)$ with probability $1/2$ each.

Thus the clusters are centered at $\pm tu$, and the direction u is the signal we would like to learn from data. Since the model is symmetric under $u \mapsto -u$, the best we can hope for is recovery of u *up to sign*.

Let $X_1, \dots, X_n$ be i.i.d. copies of X, and define the sample covariance matrix

$$\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

Throughout this problem, $\|\cdot\|_2$ denotes the Euclidean norm for vectors, and $\|\cdot\|$ denotes the operator norm for matrices.

In this problem, you may use the following two results proved earlier in the course.

- Covariance estimation for sub-Gaussian data: If $Y_1, \dots, Y_n \in \mathbb{R}^d$ are i.i.d. copies of a mean-zero random vector Y with covariance matrix Σ_Y, define

  $$\widehat{\Sigma}_Y := \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^\top.$$

  If

  $$\|\langle Y, v\rangle\|_{\psi_2} \leqslant K \|\langle Y, v\rangle\|_{L^2} \qquad \text{for all } v \in S^{d-1},$$

  then

  $$\mathbb{P}\left\{ \|\widehat{\Sigma}_Y - \Sigma_Y\| \geqslant C K^2 \left(\sqrt{\frac{d}{n}} + \frac{d}{n} \right) \|\Sigma_Y\| \right\} \leqslant 2e^{-d}.$$

- Davis–Kahan for top eigenvectors: If A, B are symmetric and $\lambda_1(A) - \lambda_2(A) = \delta > 0$, then

  $$\min_{s \in \{-1,1\}} \|v_1(A) - s\, v_1(B)\|_2 \leqslant C\, \frac{\|A - B\|}{\delta}.$$

Assume throughout that

$$\|u\|_2 = 1 \qquad \text{and} \qquad t > 0.$$

(a) **Covariance and spike structure.** This part shows that the Gaussian mixture model has a rank-one spiked covariance structure.

(a1) Show that $\mathbb{E}X = 0$ and compute the covariance matrix of X, proving that

$$\Sigma := \mathbb{E}[XX^\top] = I_d + t^2 u u^\top.$$

(2 points)

(a2) Deduce that the two largest eigenvalues of Σ are

$$\lambda_1(\Sigma) = 1 + t^2, \qquad \lambda_2(\Sigma) = 1,$$

and that the top eigenspace is $\mathrm{span}(u)$, so one may choose

$$v_1(\Sigma) = u.$$

(2 points)

(b) **Learning the signal direction from data.** We now show that the top eigenvector of the sample covariance recovers the cluster-separation direction.

(b1) Show that there exists an absolute constant K such that for every $a \in S^{d-1}$,

$$\|\langle X, a\rangle\|_{\psi_2} \leqslant K \, \|\langle X, a\rangle\|_{L^2}.$$

You may use without proof that a standard Gaussian has ψ_2-norm bounded by an absolute constant, and that a bounded random variable W satisfies

$$\|W\|_{\psi_2} \leqslant C\|W\|_\infty.$$

(3 points)

(b2) Let $v := v_1(\widehat{\Sigma})$ be a unit top eigenvector of the sample covariance matrix. Show that if

$$n \geqslant C\, d \left(\frac{1 + t^2}{t^2}\right)^2$$

for a sufficiently large absolute constant C, then the covariance-estimation bound implies

$$\|\widehat{\Sigma} - \Sigma\| \leqslant c_0 t^2$$

with probability at least $1 - 2e^{-d}$, for a sufficiently small absolute constant $c_0 > 0$. Conclude that

$$\min_{s \in \{-1,1\}} \|u - sv\|_2 \leqslant 0.1.$$

In particular, when $t \geqslant 0.1$, this condition reduces to $n \geqslant Cd$, so $n = O(d)$ unlabeled samples are enough to recover the cluster-separation direction up to sign.

(3 points)

(c) **[Bonus] Sample mean with observed labels.** Suppose now that the latent labels $\theta_1, \ldots, \theta_n$ are also observed. Define

$$\widetilde{u} := \frac{1}{nt} \sum_{i=1}^{n} \theta_i X_i.$$

Show that $\mathbb{E}\widetilde{u} = u$, and compute

$$\mathbb{E}\|\widetilde{u} - u\|_2^2.$$

Compare the scale of this error to the unlabeled PCA-type estimator from part (b2).

(2 bonus points)

2 Statistical learning via empirical processes and VC dimension

Let (X, Y) be a random pair taking values in $\mathcal{X} \times \mathcal{Y}$, and let

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

be i.i.d. copies of (X, Y). Let $\mathcal{H}$ be a hypothesis class, and let $\ell(h(X), Y) \in [0, 1]$ be a bounded loss.

For $h \in \mathcal{H}$, define the population risk

$$R(h) := \mathbb{E}[\ell(h(X), Y)]$$

and the empirical risk

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i).$$

Let

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} R_n(h), \qquad h^\star \in \arg\min_{h \in \mathcal{H}} R(h).$$

For a class $\mathcal{G}$ of real-valued functions on a sample $z_1, \ldots, z_n$, define the empirical Rademacher complexity with absolute values by

$$\widehat{\mathfrak{R}}_n(\mathcal{G}; z_1, \ldots, z_n) := \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(z_i) \right|,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher signs.

You may use the following facts from lecture and homework.

- Finite-class Rademacher bound: If $\mathcal{G}$ is finite and $0 \leqslant g(z_i) \leqslant 1$ for all $g \in \mathcal{G}$ and i, then

$$\widehat{\mathfrak{R}}_n(\mathcal{G}; z_1, \ldots, z_n) \leqslant C \sqrt{\frac{\log(e|\mathcal{G}|)}{n}}.$$

- Sauer–Shelah: If $\mathcal{H}$ is a Boolean class with $\mathrm{vc}(\mathcal{H}) = v$ and $1 \leqslant v \leqslant n$, then on any n-point sample it induces at most

$$\sum_{j=0}^{v} \binom{n}{j} \leqslant \left(\frac{en}{v}\right)^v$$

distinct labelings.

(a) **Excess-risk lemma.** Show that

$$R(\widehat{h}) - R(h^\star) \leqslant 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

(2 points)

(b) **Symmetrization.** Let

$$\mathcal{L}_\mathcal{H} := \{(x, y) \mapsto \ell(h(x), y) : \ h \in \mathcal{H}\}.$$

Prove the symmetrization bound

$$\mathbb{E} \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \leqslant 2\, \mathbb{E}\, \widehat{\mathfrak{R}}_n(\mathcal{L}_\mathcal{H}; (X_1, Y_1), \ldots, (X_n, Y_n)).$$

(3 points)

(c) **Finite hypothesis classes.** Assume $\mathcal{H}$ is finite, and the loss is bounded in $[0,1]$. Show that

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant C\sqrt{\frac{\log(e|\mathcal{H}|)}{n}}.$$

(2 points)

(d) **Boolean classification and VC dimension.** Assume now that $\mathcal{H}$ is a Boolean class, $\mathcal{Y} = \{0,1\}$, and

$$\ell(h(x), y) = \mathbf{1}\{h(x) \neq y\}$$

is the 0-1 loss. Suppose $\mathrm{vc}(\mathcal{H}) = v$ with $1 \leqslant v \leqslant n$.

Show that, on any fixed sample, the loss class induces no more distinct loss vectors than the number of labelings induced by $\mathcal{H}$, and use Sauer–Shelah to prove

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant C\sqrt{\frac{v\log(en/v)}{n}}.$$

(3 points)

(e) **[Bonus] A sample-complexity statement.** Using a high-probability version of the VC uniform convergence bound, explain why a sample size of order

$$n \gtrsim \frac{v\log(en/v) + \log(1/\delta)}{\varepsilon^2}$$

is sufficient to obtain a guarantee of the form

$$R(\widehat{h}) \leqslant R(h^\star) + \varepsilon$$

with probability at least $1 - \delta$.

(2 bonus points)

3 Nonparametric regression and the curse of dimensionality

We now study a basic idealized nonparametric regression problem.

Let X be a random point in $[0,1]^d$ with law μ, and let

$$(X_1, T(X_1)), \dots, (X_n, T(X_n))$$

be noiseless training data, where $X_1, \dots, X_n$ are i.i.d. copies of X, and $T : [0,1]^d \to [0,1]$ is an unknown target function.

For a hypothesis class $\mathcal{F} \subset \{f : [0,1]^d \to [0,1]\}$, define the population risk and empirical risk by

$$R(f) := \mathbb{E}\big[(f(X) - T(X))^2\big], \qquad R_n(f) := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - T(X_i))^2.$$

Let

$$f^\star \in \arg\min_{f \in \mathcal{F}} R(f), \qquad \widehat{f}_n \in \arg\min_{f \in \mathcal{F}} R_n(f).$$

You may use the following facts proved earlier.

- Excess-risk lemma:
$$R(\widehat{f}_n) - R(f^\star) \leqslant 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|.$$

- Empirical-process Dudley bound: If $\mathcal{G}$ is a class of functions taking values in $[0,1]$, then
$$\mathbb{E} \sup_{g \in \mathcal{G}} \big|\mu_n(g) - \mu(g)\big| \leqslant \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{G}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$

- Finite-class bound: If $\mathcal{G}$ is a finite class of functions taking values in $[0,1]$, then
$$\mathbb{E} \sup_{g \in \mathcal{G}} \big|\mu_n(g) - \mu(g)\big| \leqslant C\sqrt{\frac{\log(e|\mathcal{G}|)}{n}}.$$

For $L > 0$, define the Lipschitz class

$$\mathcal{F}_{L,d} := \big\{f : [0,1]^d \to [0,1] : \ \|f\|_{\mathrm{Lip}} \leqslant L\big\}.$$

You may use the following covering-number bounds without proof:

$$\log \mathcal{N}(\mathcal{F}_{L,1}, \|\cdot\|_\infty, \varepsilon) \leqslant C\,\frac{L}{\varepsilon}, \qquad 0 < \varepsilon \leqslant 1,$$

$$\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon) \leqslant C_d \left(\frac{L}{\varepsilon}\right)^d, \qquad 0 < \varepsilon \leqslant 1,$$

where C_d may depend on the ambient dimension d.

(a) **Loss class versus hypothesis class.** Let

$$\mathcal{L}_{\mathcal{F}} := \big\{x \mapsto (f(x) - T(x))^2 : \ f \in \mathcal{F}\big\}.$$

Show that for any $f, g \in \mathcal{F}$,

$$\|(f-T)^2 - (g-T)^2\|_\infty \leqslant 2\|f-g\|_\infty.$$

Deduce that for every $\varepsilon > 0$,

$$\mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon) \leqslant \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2).$$

(2 points)

(b) **One-dimensional Lipschitz regression.** Assume now that $d = 1$ and $\mathcal{F} = \mathcal{F}_{L,1}$. Show that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\sqrt{\frac{L}{n}}.$$

Thus, in one dimension, ERM over the class of L-Lipschitz functions achieves the same $n^{-1/2}$ scale as many parametric problems.

(3 points)

(c) **Higher-dimensional Lipschitz regression.** Now let $d \geqslant 2$ and $\mathcal{F} = \mathcal{F}_{L,d}$.

(c1) Explain why this entropy bound does not give a useful direct Dudley estimate in dimensions $d \geqslant 2$: plugging

$$\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon) \leqslant C_d(L/\varepsilon)^d$$

into the Dudley integral leads to the divergent upper-bound calculation

$$\int_0^1 \varepsilon^{-d/2}\, d\varepsilon = \infty.$$

(1 point)

(c2) Let $\mathcal{N}_\varepsilon \subset \mathcal{L}_{\mathcal{F}}$ be an ε-net of $\mathcal{L}_{\mathcal{F}}$ in $\|\cdot\|_\infty$. Show that

$$\sup_{g \in \mathcal{L}_{\mathcal{F}}} |\mu_n(g) - \mu(g)| \leqslant 2\varepsilon + \max_{h \in \mathcal{N}_\varepsilon} |\mu_n(h) - \mu(h)|.$$

(1 point)

(c3) Deduce that for every $\varepsilon \in (0,1)$,

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\left(\varepsilon + \sqrt{\frac{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon)}{n}}\right).$$

Then use part (a) and the covering-number bound for $\mathcal{F}_{L,d}$ to conclude that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C_d\left(\varepsilon + \sqrt{\frac{(L/\varepsilon)^d}{n}}\right).$$

(2 points)

(c4) Choose

$$\varepsilon \asymp L^{d/(d+2)}\, n^{-1/(d+2)}$$

and deduce the excess-risk bound

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C_d\, L^{d/(d+2)}\, n^{-1/(d+2)}.$$

Finally, explain why this is an instance of the *curse of dimensionality*.

(2 points)

(d) **[Bonus] Smoother classes help.** Suppose now that $\mathcal{F} \subset \{f : [0,1]^d \to [0,1]\}$ is a hypothesis class whose covering numbers satisfy

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leqslant A\, \varepsilon^{-p} \qquad \text{for all } 0 < \varepsilon \leqslant 1,$$

for some constants $A > 0$ and $p > 0$.

Show, using the same finite-net argument as in part (c), that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\, A^{1/(p+2)}\, n^{-1/(p+2)}.$$

Now suppose moreover that $\mathcal{F}$ is a class of s-smooth functions for which $p = d/s$. Treating A as a constant, and taking $p = d/s$, deduce the heuristic rate

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \lesssim n^{-s/(2s+d)}.$$

Explain briefly why increasing the smoothness s improves the rate.

(2 bonus points)