Algorithmic Statistics Problem Set 1

Problem 1 - Uniformity Testing

In this exercise you will complete a missing step from the proof of the uniformity testing lower bound from lecture 1.

Using facts from the lecture notes about Poissonization and χ^2 divergence, finish the proof of Paninski's sample complexity lower bound (poissonized variant) by proving the poissonized version of Lemma 5.2 from the lecture 1 notes. You may look up and use any other facts about Poisson random variables.

Problem 2: Convergence of Linear Regression

In this problem we analyze one of the most fundamental problems in learning theory: linear regression. Before delving into the analysis it is worth noting a few of the differences between it and many of the other analyses of linear regression.

1. *Agnostic setting.* We do not assume that $y = \langle x, \beta_{\text{gt}} \rangle + \text{noise}$ for some ground truth β_{gt}. Instead, we simply define the *best linear predictor* $\beta^\star$ by population least-squares:

   $$\beta^\star = \arg\min_\beta \ \mathbb{E}\left[(y - \langle x, \beta \rangle)^2\right].$$

   This makes the results applicable to a wider variety of settings.

2. *Parameter convergence vs. risk convergence.* Here we study convergence of the parameters both in ℓ_2 norm (i.e. $\left\|\hat{\beta} - \beta^\star\right\|_2 \le \varepsilon$), and in *prediction risk*:

   $$\mathbb{E}\left[(y - \langle x, \hat{\beta} \rangle)^2\right] \ \le \ \min_\beta \mathbb{E}\left[(y - \langle x, \beta \rangle)^2\right] \ + \ (\text{generalization error}).$$

3. *Fast vs. slow rates.* By exploiting matrix concentration and boundedness assumptions, we obtain sample complexity

   $$n = \tilde{O}\left(\tfrac{d}{\varepsilon^2}\right),$$

   sometimes called a *fast rate*. In contrast, "black-box" VC-dimension style theory typically yields only

   $$n = \tilde{O}\left(\tfrac{d}{\varepsilon^4}\right),$$

   a *slow rate*.

4. *Bounded Covariates and Residuals* One somewhat unorthodox assumption we make is that the covariates x_i and the labels y_i are bounded almost-surely (i.e., with probability 1), where many analyses would only assume something like subgaussian covariates and labels.

   We make this assumption to simplify the use of matrix Bernstein (see below), but with a little more work, the same techniques can be used with just a tail bound on these quantities.

Setting Let $(x_i, y_i)_{i=1}^n$ be i.i.d., with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, satisfying

- (*isotropy*) $\mathbb{E}[x] = 0$ and $\Sigma := \mathbb{E}\left[xx^\top\right] = I_d$,
- (*bounded norm*) $\|x\|_2 \le C\sqrt{d}$ almost surely, for an absolute constant C,
- (*bounded labels*) $|y| \le 1$ almost surely.

Definitions Define the *population* quantities

$$\beta^\star := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left\{ \mathbb{E}\left[(y - \langle \beta, x \rangle)^2\right] \right\} = \mathbb{E}\left[\Sigma^{-1} x y\right] =: g, \qquad \Sigma = I_d,$$

and the *empirical* quantities

$$\widehat{\Sigma} = \tfrac{1}{n} \sum_{i=1}^n x_i x_i^\top, \qquad \widehat{g} = \tfrac{1}{n} \sum_{i=1}^n x_i y_i, \qquad \widehat{\beta} = \widehat{\Sigma}^{-1} \widehat{g}.$$

Matrix Bernstein Inequality You may use the following without proof.

Lemma 1 (Matrix Bernstein Inequality)**.** *Let $Z_1, \dots, Z_n$ be independent mean-zero, self-adjoint random matrices in $\mathbb{R}^{d\times d}$ with $\|Z_i\| \le L$ almost surely, and set $v = \left\|\sum_i \mathbb{E}\left[Z_i^2\right]\right\|$. Then for all $t \ge 0$,*

$$\Pr\left(\left\|\sum_{i=1}^n Z_i\right\| \ge t\right) \;\le\; 2d\,\exp\left(-\frac{t^2}{2v + \frac{2}{3}Lt}\right).$$

Goal Our goal will be to prove a bound on the convergence of linear regression.

Claim 2. *Given $n = \tilde{O}\left(\frac{d}{\varepsilon^2}\right)$, with probability $1 - o(1)$, the empirical regression solution $\widehat{\beta}$ is close to the population regression solution $\beta^\star$, up to an L_2 error of $\left\|\widehat{\beta} - \beta^\star\right\| \le \varepsilon$.*

—

2A. Concentration of the covariance $\widehat{\Sigma}$

Let $Z_i := x_i x_i^\top - \Sigma = x_i x_i^\top - I_d$. Note that $\widehat{\Sigma} - I_d = \frac{1}{n}\sum_{i=1}^n Z_i$.

(i) **Mean Zero.** Show that $\mathbb{E}\left[Z_i\right] = 0$.

(ii) **Operator-norm Bound on Z_i.** Recall $\|A\| := \sup_{\|u\|=1} \|Au\|_2$.
Prove that $\|Z_i\| = O(d)$ almost surely.

(iii) **Variance Property.** Show that

$$Z_i^2 = (x_i x_i^\top - I_d)^2 = \left(\|x_i\|_2^2 - 2\right) x_i x_i^\top \;+\; I_d,$$

and conclude that

$$\left\|\mathbb{E}\left[Z_i^2\right]\right\| = O(d), \qquad \left\|\sum_{i=1}^n \mathbb{E}\left[Z_i^2\right]\right\| = O(nd).$$

(iv) **Matrix Bernstein Inequality.** Apply Lemma 1 to $\sum_{i=1}^n Z_i$ and show that with probability at least $1 - \delta$,

$$\left\|\widehat{\Sigma} - I_d\right\| \;\le\; O\left(\sqrt{\frac{d\log\frac{2d}{\delta}}{n}} + \frac{d\log\frac{2d}{\delta}}{n}\right).$$

(v) **Stability of the inverse.** Show that if $\left\|\widehat{\Sigma} - I_d\right\| \le \eta < 1$, then

$$\left\|\widehat{\Sigma}^{-1} - I_d\right\| \;\le\; \frac{\eta}{1-\eta}.$$

—

2B. Concentration of the linear term $\widehat{g}$

We need a high-probability bound on $\|\widehat{g} - g\|_2$.

To this end, define $v_i = x_i y_i - g$ and

$$V_i = \begin{pmatrix} 0 & v_i \\ v_i^\top & 0 \end{pmatrix}.$$

(i) **Mean Zero.** Show that $\mathbb{E}[V_i] = 0$.

(ii) **Operator-norm Bound.** Show that $\|V_i\| = O(\sqrt{d})$ almost surely.

(iii) **Variance Property.** Show that

$$\left\|\mathbb{E}\left[V_i^2\right]\right\| = O(d), \qquad \left\|\sum_{i=1}^{n}\mathbb{E}\left[V_i^2\right]\right\| = O(nd).$$

(iv) **Matrix Bernstein.** Conclude that with probability at least $1-\delta$

$$\|\widehat{g} - g\|_2 = O\left(\sqrt{\tfrac{d\log(d/\delta)}{n}} + \tfrac{\sqrt{d}\log(d/\delta)}{n}\right).$$

—

2C. Concluding regression convergence

Use the previous parts of the question to prove Claim 2.

(i) **Key identity.** Prove

$$\widehat{\beta} - \beta^\star \;=\; (\widehat{\Sigma}^{-1} - I_d)\, g \;+\; \widehat{\Sigma}^{-1}(\widehat{g} - g).$$

(ii) **Bounding each term.** Use the bounds from 2A and 2B.

(iii) **Final conclusion.** Deduce that for $n = \tilde{O}(d/\varepsilon^2)$, with high probability $\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \varepsilon$.

—

2D. Extensions: Generalization Errors in General Settings

Explore how the conclusions change under weaker assumptions.

(i) **Non-isotropic distributions.** Suppose instead $\Sigma = \mathrm{Cov}(x)$ is not the identity, but satisfies $x^\top \Sigma^{-1} x = O(d)$ almost surely. Show a similar bound holds on

$$\left\|\Sigma^{1/2}(\widehat{\beta} - \beta^\star)\right\|.$$

(ii) **Bounded residuals rather than bounded labels.** Suppose instead of the bounded label assumption, we are told that some "ground-truth" β_{gt} such that

$$y_i = \langle x_i, \beta_{\mathrm{gt}}\rangle + e_i, \qquad e_i \in [-1, 1].$$

Show that the same convergence analysis applies even though the labels are not necessarily bounded (*hint*: analyze the optimization problem $\gamma^* = \mathrm{argmin}_{\gamma\in\mathbb{R}^d}\left\{\mathbb{E}\left[\left(y - \langle x, (\gamma - \beta_{\mathrm{gt}})^2\rangle\right)\right]\right\}$).

(iii) **Bounding the generalization error.** Suppose $|y - \langle\beta_{\mathrm{gt}}, x\rangle| \le \eta$ almost surely. Bound the excess risk

$$\mathbb{E}\left[\left(y - \langle\widehat{\beta}, x\rangle\right)^2\right] - \mathbb{E}\left[(y - \langle\beta_{\mathrm{gt}}, x\rangle)^2\right].$$

(*hint*: show that $\mathbb{E}[x\,(y - \langle\beta_{\mathrm{gt}}, x\rangle)]$, and then express the excess risk in terms of $\delta \coloneqq \beta_{\mathrm{gt}} - \beta$)

Problem 3: Learning Halfspaces

In this problem we will build towards the main theorem of Kalai–Klivans–Mansour–Servedio: *halfspaces can be weakly agnostically learned under the uniform distribution.*

Background

- **Why agnostic learning?** Real-world data may be noisy or not generated exactly by our model class. Agnostic learning relaxes the assumptions: the data distribution is arbitrary, and we only require our learner to perform nearly as well as the best model in hindsight.

  In other words, we still assume that our training samples (x, y) are iid, but we no longer assume that the distribution behaves like a model plus random noise.

- **Why halfspaces?** Halfspaces (linear threshold functions) are fundamental classifiers of the form

$$h(x) = \operatorname{sign}(w \cdot x - \theta), \qquad x \in \{\pm 1\}^n .$$

- **Main result (informal).** Kalai–Klivans–Mansour–Servedio (2005) show that halfspaces are agnostically learnable to error $\mathrm{opt} + \varepsilon$. In this problem, you will prove a weaker guarantee: there exists an efficient *weak agnostic learner* for halfspaces under the uniform distribution, achieving error

$$O\left(\mathrm{opt} + \varepsilon\right).$$

- **Setup.** We are given training data $\{(x_i, y_i)\}_{i=1}^m$, with $x_i \in \{\pm 1\}^n$ features and $y_i \in \{\pm 1\}$ labels. The key property we will exploit is that every halfspace can be well-approximated by a low-degree Fourier polynomial.

3A. Fourier characters on the hypercube

(i) Definition and orthogonality. For a subset $S \subseteq [n]$, define the Fourier character

$$\chi_S(x) = \prod_{i \in S} x_i.$$

Prove that for subsets $S, T \subseteq [n]$:

$$\mathop{\mathbb{E}}_{x \sim \{\pm 1\}^n} \left[\chi_S(x)\chi_T(x)\right] = \begin{cases} 1 & S = T, \\ 0 & S \neq T. \end{cases}$$

Hint: Show that $\chi_S \cdot \chi_T = \chi_{S \,\Delta\, T}$, where Δ is symmetric difference.

(ii) Low-degree feature map. Define the feature map $\phi : \{\pm 1\}^n \to \mathbb{R}^N$ by

$$\phi(x) = (\chi_S(x))_{S : 1 \le |S| \le d},$$

i.e. all non-empty characters of degree at most d. Show that under the uniform distribution:

- $\mathbb{E}_{x \sim \{\pm 1\}^n}\left[\phi(x)\right] = 0$,
- $\mathbb{E}_{x \sim \{\pm 1\}^n}\left[\phi(x)\phi(x)^\top\right] = I_N$.

Thus, ϕ is isotropic (see Problem 2).

3B. Learning low-degree Boolean functions with noise

We now show that if the true function is close to a low-degree polynomial, then regression against noisy samples still recovers a good hypothesis.

Definition 1 (Low-degree Approximability)**.** *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be a Boolean function. We say that f can be ε-approximated by a degree d polynomial in ℓ_2 norm if there exists $p^\star$ in the span of $\{\chi_S : |S| \le d\}$ such that*

$$\mathop{\mathbb{E}}_{x\sim\{\pm 1\}^n}\left[(f(x) - p^\star(x))^2\right] \le \varepsilon.$$

Claim 3. *Let $f, g : \{\pm 1\}^n \to \{\pm 1\}$ be Boolean functions, where f represents the optimal target and g represents noisy training labels. Suppose:*

1. *f is ε-approximated by a degree d polynomial.*
2. *(*Bounded noise*)*

$$\Pr_{x\sim\{\pm 1\}^n}[f(x) \ne g(x)] \le \eta.$$

Let

$$p \;=\; \arg\min_{q\in \mathrm{span}\{\chi_S : |S|\le d\}} \mathop{\mathbb{E}}_{x\sim\{\pm 1\}^n}\left[(g(x) - q(x))^2\right].$$

Then

$$\Pr_{x\sim\{\pm 1\}^n}[\,|p(x) - g(x)| > 1/2] = O(\eta + \varepsilon).$$

(i) Almost-triangle inequality. Show that for any reals a, b, c,

$$(a-c)^2 \le 2\big((a-b)^2 + (b-c)^2\big).$$

(ii) $p^\star$ has small loss against g. Use the inequality from part (i) to prove:

$$\mathop{\mathbb{E}}_x\left[(p^\star(x) - g(x))^2\right] \le 2\mathop{\mathbb{E}}_x\left[(f(x) - g(x))^2\right] + 2\varepsilon.$$

Conclude that

$$\mathop{\mathbb{E}}_x\left[(p^\star(x) - g(x))^2\right] \le 8\eta + 2\varepsilon.$$

Why does this imply that the regression minimizer p must also have loss at most $8\eta + 2\varepsilon$?

(iii) Conclude the claim. Explain why the bound on squared error implies that p misclassifies at most an $O(\eta + \varepsilon)$ fraction of inputs. *Hint:* If $|p(x) - g(x)| > 1/2$, then the squared error is at least $1/4$.

3C. Agnostically learning halfspaces

We now combine everything.

It is a known fact (you may assume without proof) that every halfspace $h(x) = \mathrm{sign}(w \cdot x - \theta)$ can be ε-approximated in L_2 under the uniform distribution by a polynomial of degree $\mathrm{poly}(1/\varepsilon)$. Thus, halfspaces fall into the approximately low-degree setting of Problem 3B.

(i) Finite-sample regression. Using Problem 2 and part 3A, show that with $m = \mathrm{poly}(N, n)$ samples, regression recovers a hypothesis $\hat{p}$ whose coefficient vector in $\mathbb{R}^N$ satisfies

$$\|\hat{p} - p\|_2 \;<\; \tfrac{1}{2Nn}.$$

(i.e., bound the ℓ_2 norm of the $\mathbb{R}^N$ vector whose Sth entry is the difference in the coefficient of χ_S in $\hat{p}$ and in p)

(ii) From coefficient closeness to classification error. Show that the ℓ_2-closeness of coefficient vectors implies ℓ_∞ closeness of predictions

$$\max_{x \in \{\pm 1\}^n} \{|p(x) - \hat{p}(x)|\} = \|\hat{p} - p\|_\infty < \tfrac{1}{2}.$$

Hint: Use Cauchy–Schwarz together with the triangle inequality.

Finally, use this and the result of Problem 3B to conclude that $\hat{p}$ misclassifies at most $O(\eta + \varepsilon)$ fraction of the domain. Thus, there is a polynomial-time weak agnostic learner for halfspaces under the uniform distribution.