ESE 6180: Learning, Dynamics and Control **Fall 2024**

Homework 2

Assigned: 9/19/2024 *Due: 10/16/2024*

Homework must be LaTeX'd or it will not be graded.

Grading: Each problem will be graded on a scale of 0-4. If you get 80% of the problem or more correct, and make an honest attempt at the rest, you will get 4/4. If you get 60% of the problem or more correct, you will get 3/4, etc.

Canvas: Please submit your HW as a single pdf file, with pages correctly tagged to go with each problem.

Working in groups: You are allowed to, and in fact encouraged, to discuss and work on problems with your classmates. However, each student must write up their own homework independently. Further, please make note of your collaborators in the designated spot in the homework template.

Citing references: If you referred to solutions found in published material (papers, textbooks, websites, etc.), you must cite these in your homework solutions. It is ok to use proofs that you find online for guidance, but you should indicate where and how you did so, and you should always make a first attempt at the answer on your own. Importantly, even if you are following the guidance of a proof from a paper, you must be sure to fully explain all steps, as well as fill in any missing steps.

Useful inequalities: This cheat sheet may come in handy throughout the course.

1 Linear Regression

This exercise asks you to prove an analogue of the concentration inequality for the least squares estimator without imposing sub-Gaussianity of the $W_i X_i$ and X_i^2. Assume that both the variables $X_i \in \text{subG}(\sigma_X^2), i = 1, \dots, n$ and the $W_i \in \text{subG}(\sigma_W^2), i = 1, \dots, n$, are drawn iid and that $W_{1:n}$ is mean zero. Recall also that the least squares estimator for $\theta_\star$ is given as:

$$\widehat{\theta} - \theta_\star = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n X_i^2}$$

whenever $X_i \neq 0$ for at least index one $i \in [n]$.

(a) Show there exist universal positive constant $c, c' > 0$ such that for $\delta \in (0, 1)$ it holds with probability at least $1 - \delta$ that:

$$\frac{1}{n} \sum_{i=1}^n W_i X_i \leq \sqrt{\frac{c \sigma_X^2 \sigma_W^2 \log(1/\delta)}{n}} + \frac{c'(\sigma_X(1 + \sigma_W)) \log(1/\delta)}{n}. \tag{1}$$

(b) Show that there exist universal positive constants $c, c' > 0$ such that with probability at least $1 - \delta$

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \geq c \mathbf{E} X^2 \tag{2}$$

as long as $n \geq \frac{c' \sigma_X^2 \log(1/\delta)}{\mathbf{E} X^2}$.

(c) Conclude that there exist universal positive constants c, c' such that with probability at least $1 - 3\delta$ we have that

$$|\widehat{\theta} - \theta_\star| \le \sqrt{\frac{c(\sigma_X \sigma_W)^2 \log(1/\delta)}{n(\mathbf{E}X^2)^2}} \tag{3}$$

as long as $n \ge c' \left(\frac{(1+\sigma_W)^2}{\sigma_W^2} \vee \frac{\sigma_X^2}{\mathbf{E}X^2} \right) \log(1/\delta)$.

(d) Note that (3) is still unsatisfactory in that the leading term depends on $\sigma_X \sigma_W$ which is qualitatively larger than the variance of XW. Show that if X and W are bounded by B_X and B_W then $\sigma_X \sigma_W$ can be replaced by $\mathbf{V}(XY)$ in (3) at the cost of inflating the burn-in ($n \ge ...$).

2 Concentration and Covering

(a) Let R be a Rademacher random variable, i.e., R takes values -1 and 1 with probability $1/2$ each. Show that $R \in \mathrm{subG}(1)$.

(b) This exercise walks you through an alternative approach for showing that the square of a sub-Gaussian squares has sub-exponential tails. Let $G \sim N(0,1)$ and fix a centered $X \in \mathrm{subG}(\sigma^2)$ coming from two independent sources of randomness.

(i) Show that for every $\lambda \in \mathbb{R}$ satisfying $|\lambda| < 1/2$ we have that:

$$\mathbf{E} \exp\left(\lambda G^2\right) = \frac{1}{\sqrt{1-2\lambda}}. \tag{4}$$

(ii) Show that for every $\lambda \in \mathbb{R}$ satisfying $|\lambda| < 1/2\sigma^2$ we have that:

$$\mathbf{E} \exp\left(\lambda X^2\right) \le \frac{1}{\sqrt{1-2\sigma^2\lambda}}. \tag{5}$$

Hint: What is the "partial MGF" of XG, integrated only with respect to X?

(iii) Conclude that for every $\lambda \in \mathbb{R}$ satisfying $0 \le \lambda < 1/4\sigma^2$ we have that:

$$\mathbf{E} \exp\left(\lambda(X^2 - \sigma^2)\right) \le \exp\left(2\lambda^2\sigma^4\right). \tag{6}$$

Compare to the result in the lecture notes. How does the result differ?

(c) Fix $\varepsilon \in (0, 1/2)$, let $M \in \mathbb{R}^{d \times d'}$ and let $\mathcal{N}, \mathcal{M}$ be ε-nets of $\mathbb{S}^{d-1}$ and $\mathbb{S}^{d'-1}$.

(i) Show that

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Mx, y \rangle \le \|M\|_{\mathsf{op}} \le \frac{1}{1-2\varepsilon} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Mx, y \rangle. \tag{7}$$

(ii) Moreover if $d = d'$ and M is symmetric show that the result can be simplified:

$$\sup_{x \in \mathcal{N}} \langle Mx, x \rangle \le \|M\|_{\mathsf{op}} \le \frac{1}{1-2\varepsilon} \sup_{x \in \mathcal{N}} \langle Mx, x \rangle. \tag{8}$$

Hint: Proceed similarly to the proof of the 2-norm covering bound in the lecture notes and use the identity $\langle Mx, y \rangle - \langle Mx', y' \rangle = \langle Mx, y - y' \rangle - \langle M(x' - x), y' \rangle$.

(d) Let $M \in \mathbb{R}^{d \times d}$ be a random matrix with independent mean zero σ^2-sub-Gaussian entries.

(i) Prove that there exists a universal positive constant $c > 0$ such that with probability $1 - \delta$:

$$\|M\|_{\mathsf{op}} \leq c\sigma(\sqrt{d} + \sqrt{\log(1/\delta)}).$$

Hint: For every $u, v \in \mathbb{S}^{d-1}$, $\sum_{i,j}^{d} M_{ij} u_i v_j$ is a sum of independent sub-Gaussian random variables.

(ii) Prove that there exists a universal positive constant $c > 0$ such that:

$$\mathbf{E}\|M\|_{\mathsf{op}} \leq c\sigma\sqrt{d}.$$

(iii) The bound above is not improvable in general. Show that if the entries are Gaussian with unit variances that for sufficiently large d and some universal constant $c > 0$:

$$\mathbf{E}\|M\|_{\mathsf{op}} \geq c\sqrt{d}.$$

3 Linear Dynamical Systems

This exercise asks you to run a few basic numerical experiments on linear regression in a first order auto-regression in dimension $d_{\mathsf{X}} \in \mathbb{N}$. Namely, suppose that you are given access to $m \in \mathbb{N}$ trajectories of length $T \in \mathbb{N}$ of the process $X^{(j)}_{1:T+1}$, $j \in [m]$ specified as follows. Let for each j the process, $W^{(j)}_{1:T+1}$ be iid standard normal and let the $X^{(j)}_{1:T+1}$ satisfy:

$$X^{(j)}_{k+1} = A_\star X^{(j)}_k + W^{(j)}_{k+1} \qquad X^{(j)}_1 = W^{(j)}_1 \qquad k = 1, \ldots, T, \qquad j \in [m]. \tag{9}$$

where $A_\star \in \mathbb{R}^{d_{\mathsf{X}} \times d_{\mathsf{X}}}$ is a fixed matrix (shared across the different trajectories—the interpretation is that these are different rollouts from the same dynamical system with randomized initial conditions). Notice that in total you have $n \triangleq m \times T$ data points at your disposal but that there are correlations across the time axis.

We now specify $A_\star$ as a single Jordan block with eigenvalue $\lambda \in \mathbb{R}$ we let

$$A_\star = \lambda I_{d_{\mathsf{X}}} + \mathbf{Z}^{\mathsf{T}} I_{d_{\mathsf{X}}} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & \ldots & 0 \\ 0 & \lambda & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & \cdots & \cdots & 1 & 0 \\ \vdots & \cdots & \cdots & \cdots & \lambda & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda \end{bmatrix} \tag{10}$$

where $\mathbf{Z}^{\mathsf{T}}$ is the upshift operator. You may pick $d_{\mathsf{X}} \geq 2$ freely in the sequel (try a few different ones).

(a) In the first few classes we only studied iid linear regression, and being a very cautious student, you decide to only run linear regression on something you know you can prove a guarantee on. Namely, we throw away all samples but the last from each trajectory. Let $Y_j = X^{(j)}_{T+1}$ and $X^{\text{last}}_j = X^{(j)}_T$ and form the least squares estimate

$$\widehat{A}^{\text{last}} = \left(\sum_{i=1}^{m} Y_i (X^{\text{last}}_i)^{\mathsf{T}} \right) \left(\sum_{i=1}^{m} X^{\text{last}}_i (X^{\text{last}}_i)^{\mathsf{T}} \right)^{-1}. \tag{11}$$

Verify that this estimator is consistent by plotting by the performance for a varying range of $T \in \mathbb{N}$ and $m \gg d_{\mathsf{X}}$ and $\lambda \in \{0.5, 0.7, 0.9, 1, 1.1, 1.3\}$. That is, plot the error $\widehat{A}^{\text{last}} - A_\star$ computed in your favorite norm.

(b) Define now instead $Y^{(j)}_{1:k} = X^{(j)}_{2:k+1}$ as the "observations" for trajectory j, i.e. so that $Y^{(j)}_k = A_\star X^{(j)}_k + W^{(j)}_{k+1}$. Finally, we stack all the data together ($n = mT$ total data points):

$$X_{1:n} = \begin{bmatrix} X^{(1)}_{1:T} \\ \vdots \\ X^{(m)}_{1:T} \end{bmatrix}, \quad Y_{1:n} = \begin{bmatrix} X^{(1)}_{2:T+1} \\ \vdots \\ X^{(m)}_{2:T+1} \end{bmatrix} \quad \text{and} \quad W_{1:n} = \begin{bmatrix} W^{(1)}_{2:T+1} \\ \vdots \\ W^{(m)}_{2:T+1} \end{bmatrix} \tag{12}$$

and form the single least squares estimate:

$$\widehat{A} = \left(\sum_{i=1}^{n} Y_i X_i^\mathsf{T} \right) \left(\sum_{i=1}^{n} X_i X_i^\mathsf{T} \right)^{-1}. \tag{13}$$

(i) Plot the error $\widehat{A} - A_\star$ and compare this to the error $\widehat{A}^{\text{last}} - A_\star$ using the same data. Is there a speed-up from using correlated data? Vary $\lambda \in \{0.5, 0.7, 0.9, 1, 1.1, 1.3\}$ and let $m \gg d_\mathsf{X}$.

(ii) Let us now consider the case where m is relatively small compared to d_X. Plot the errors $\widehat{A} - A_\star$ and $\widehat{A}^{\text{last}} - A_\star$ and again vary $\lambda \in \{0.5, 0.7, 0.9, 1, 1.1, 1.3\}$, but this time let m either be of comparable or smaller order of magnitude than d_X—try in particular plot also the extreme case $m = 1$ corresponding to data from a single correlated trajectory.

(iii) Comment on your findings. What is the interplay between the eigenvalue λ and the number of trajectories required to learn?

4 The Johnson-Lindenstrauss Lemma

The use case of concentration inequalities (and high-dimensional probability) goes far beyond proving guarantees for linear regression. One particularly useful idea is that of *sketching* and randomized algorithms.

Fix a sequence of vectors $x_{1:n}$, with entries in $\mathbb{R}^d$. We would like to find a lower dimensional sequence $\tilde{x}_{1:n}$ with entries in $\mathbb{R}^k$ ($k < d$) such that as much as possible of the spectral information of $x_{1:n}$ is preserved. The simplest form of such a reduction is a projection. The Johnson-Lindenstrauss Lemma shows that if we allow for randomized projections, it suffices to take $k \asymp \log n$.

(a) Show that if $n < d$ then there exists a projection $P \in \mathbb{R}^{n \times d}$ such that $\|Px_i - Px_j\| = \|x_i - x_j\|$ for all pairs i, j.

(b) Let now $Y_{1:d} \sim N(0, I_d)$ and let $E_k \in \mathbb{R}^{d \times k}$ be the projection onto the first k coordinates (so that $E_k z_{1:d} = z_{1:k}$ for any sequence $z_{1:d}$ taking values in $\mathbb{R}$). Let us define $Z = \frac{1}{\|Y_{1:d}\|} E_k Y_{1:d} = \frac{1}{\|Y_{1:d}\|} Y_{1:k}$ and set also $L = \|Z\|^2$.

(i) Prove that $\mathbf{E}L = \frac{k}{d}$. Hint: what is the distribution of $Y_{1:d}/\|Y_{1:d}\|$?

(ii) Fix $\beta \in (0, \infty)$ and show that for every t satisfying $1 - 2t(k\beta - d) > 0$ and $1 - 2tk\beta > 0$ we have that

$$\mathbf{P}\left(\sum_{i=1}^{k} Y_i^2 \leq \beta \frac{k}{d} \sum_{i=1}^{d} Y_i^2 \right) \leq \frac{1}{(1 - 2tk\beta)^{-(d-k)/2}} \times \frac{1}{(1 - 2t(k\beta - d))^{-k/2}} \tag{14}$$

(iii) Show that for $\beta < 1$:

$$\mathbf{P}\left(L \leq \frac{\beta k}{d} \right) \leq \exp\left(\frac{k(1 - \beta + \log \beta)}{2} \right) \tag{15}$$

(iv) Similarly for $\beta > 1$:

$$\mathbf{P}\left(L \geq \frac{\beta k}{d} \right) \leq \exp\left(\frac{k(1 - \beta + \log \beta)}{2} \right) \tag{16}$$

(v) Establish the elementary inequalities

$$x \in (0,1) \Rightarrow \log(1-x) \leq -x - \frac{x^2}{2} \tag{17}$$

and

$$x \in (-1,\infty) \Rightarrow \log(1+x) \leq x - \frac{x^2}{2} + \frac{x^3}{3} \tag{18}$$

(c) Show that there exists a randomized projection operator $Q_{\mathrm{rand}} \in \mathbb{R}^{k \times d}$ with $\mathbf{E}\|Q_{\mathrm{rand}}x\|^2 = \frac{k}{d}\|x\|^2$ for every $x \in \mathbb{R}^d$. Hint: show that he length of a unit vector in $\mathbb{R}^d$ when it is projected onto a random k-dimensional subspace has the same distribution as the length of a random unit vector projected down onto a fixed k-dimensional subspace.

(d) Show that there exists a universal positive constant $c > 0$ such that for every $\varepsilon > 0$ and $x \in \mathbb{R}^d$ we have that

$$\frac{k(1-\varepsilon)}{d}\|x\|^2 \leq \|Q_{\mathrm{rand}}x\|^2 \leq \frac{k(1+\varepsilon)}{d}\|x\|^2 \tag{19}$$

with probability at least $1 - 2\exp\left(-ck\varepsilon^2\right)$

(e) Show that there exist universal positive constant $c, c' >$ such that if $k \geq c'\varepsilon^{-2}\log n$ then there exists a randomized projection matrix $P \in \mathbb{R}^{k \times d}$ onto a k-dimensional subspace such that

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|P(x_i - x_j)\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2 \tag{20}$$

simultaneously for every index i, j and with probability at least $1 - 2\exp\left(-ck\varepsilon^2\right)$ (over the randomness in P).