SDS 387, Fall 2024
Homework 4

Due November 14, by midnight on Canvas.

1. (For this problem you may want to consult Chapter 5 of Bach's book). Let Φ be a $n \times d$ marix and Y an n-dimensional vector. We want to use gradient descent to solve the least squares problem and find a solution

$$\widehat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{2} \|Y - \Phi\beta\|^2.$$

When Φ is of full column rank, we saw in class that the gradient descent iterations converge to the OLS estimator from any starting point. Assume instead that $\mathrm{rank}(\Phi) = n < d$. Show that gradient descent initialized at 0 will converge to the min-norm solutions

$$\widehat{\beta}_{\mathrm{min-norm}} = \Phi^{+}Y = (\Phi^\top \Phi)^{+} \Phi^\top Y = \mathrm{argmin}\left\{\|\beta\| \colon Y = \Phi\beta\right\}.$$

Bonus. Show that the same result is true when gradient descent is initialized at any point in the row span of Φ.

2. Assume a linear regression model with fixed (i.e., deterministic) covariate matrix Φ and denote with $\beta^\star \in \mathbb{R}^d$ the true regression parameters. Let $\widehat{\beta}$ be the OLS. Let $c \in \mathbb{R}^d$ be a non-zero vector. Show that the best linear unbiased estimator of $c^\top \beta^\star$ is $c^\top \widehat{\beta}$. (To be clear, you have to show that, among all estimators of the form $a^\top Y$, where Y is the vector of responses, such that $\mathbb{E}[a^\top Y] = c^\top \beta^\star$, $c^\top \widehat{\beta}$ has the smallest variance).

3. Consider the min-norm least squares estimator $\widehat{\beta}_{\mathrm{MN}} = \Phi^{+}Y$, where $Y \in \mathbb{R}^n$ and Φ is a $n \times d$ design matrix of rank r. You may want to think of R as $d < n$ but this is not necessary; in particular, the results below works also when $r = n < d$. Let U be the $n \times r$ matrix containing the left singular vectors of Φ (spanning the column space of Φ, an r-dimensional linear subspace of $\mathbb{R}^n$). Using the expression of $\widehat{\beta}_{\mathrm{MN}}$ show that the vector of fitted values is still the orthogonal projection of Y onto the column space of Φ, i.e.

$$\Phi \widehat{\beta}_{\mathrm{MN}} = \sum_{i=1}^{r} u_i \langle u_i, Y \rangle. \tag{1}$$

Now consider instead the ridge estimator

$$\widehat{\beta}_\lambda = \mathrm{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \|Y - \Phi\beta\|^2 + \lambda \|\beta\|^2,$$

where λ is a positive parameter. As we know, $\widehat{\beta}_\lambda$ is unique, even if Φ is rank deficient. Show that

$$\Phi \widehat{\beta}_\lambda = \sum_{i=1}^{r} u_i \langle u_i, Y \rangle \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \tag{2}$$

where σ_i is the ith singular value of Φ. This shows that

$$\widehat{\beta}_{\mathrm{MN}} = \lim_{\lambda \downarrow 0} \widehat{\beta}_\lambda.$$

Compare (1) and (2) and intepret.

4. Recall that Loewner order is a partial order on the set of positive semidefinite (psd) matrices. Give an example of two psd matrices A and B such that $A \npreceq B$ and $A \nsucceq B$.

5. Exercise 3.2 on page 54 of Bach's book.

6. Exercise 3.5 on page 57 of Bach's book.

7. Exercise 3.6 on page 60 of Bach's book.

8. **Quadratic forms.**

   (a) Let X be a d dimensional random vector with mean μ and covariance matrix Σ. Let A a $d \times d$ symmetric matrix. Show that $\mathbb{E}[X^\top A X] = \mathrm{tr}(A\Sigma) + \mu^\top A \mu$.

   (b) If $X \sim N_d(\mu, \Sigma)$ and $\mathcal{P}$ is an orthogonal projection matrix in $\mathbb{R}^d$, show that $X^\top \mathcal{P} X \sim \chi_r^2(\mu^\top \mathcal{P} \mu / 2)$, where r is the rank of $\mathcal{P}$. You can use these facts: if $X \sim N_d(\mu, I_d)$ then $\|X\|^2 \sim \chi_d(\|\mu\|^2/2)$ and, for any $r \times d$ matrix A with $\mathrm{rank}(A) = r \leq d$, $AZ \sim N_r(A\mu, A\Sigma A^\top)$.