Homework 3

Advanced Topics in Statistical Learning, Spring 2023

Due Friday March 24 at 5pm

1 Carathéodory's view on sparsity of lasso solutions [10 points]

In this exercise, we will prove the fact we cited in lecture about sparsity of lasso solutions, by invoking Caratheodory's theorem. Let $Y \in \mathbb{R}^n$ be a response vector, $X \in \mathbb{R}^{n \times d}$ be a predictor matrix, and consider the lasso estimator defined by solving

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

for a tuning parameter $\lambda > 0$.

(a) Let $\hat{\beta}$ be any solution to the lasso problem. Let $\hat{\alpha} = \hat{\beta}/\|\hat{\beta}\|_1$. Prove that $X\hat{\alpha}$ lies in the convex hull of the vectors [2 pts]

$$\{\pm X_j\}_{j=1}^d.$$

Note: here $X_j \in \mathbb{R}^n$ denotes the j^{th} column of X.

(b) Recall that Carathéodory's theorem states the following: given any set $C \subseteq \mathbb{R}^k$, every element in its convex hull $\text{conv}(C)$ can be represented as a convex combination of $k+1$ elements of C.

Use this theorem and part (a) to prove that there exists a lasso solution $\tilde{\beta}$ with at most $n+1$ nonzero coefficients. [2 pts]

Hint: start with a generic solution $\hat{\beta}$, and use Carathéodory's theorem to construct a coefficient vector $\tilde{\beta}$ such that (i) the fit is the same, $X\tilde{\beta} = X\hat{\beta}$; (ii) the penalty is at worst the same, $\|\tilde{\beta}\|_1 \leq \|\hat{\beta}\|_1$; and (iii) $X\tilde{\beta}$ is a nonnegative linear combination of at most $n+1$ of $\pm X_j$, $j = 1, \ldots, d$.

(c) Now, assuming $\lambda > 0$, use the subgradient optimality condition for the lasso problem to prove that the fit $X\tilde{\beta}$ from part (b) is supported on a subset of [3 pts]

$$\{\pm X_j\}_{j=1}^d$$

that has affine dimension at most $n-1$.

Hint: this is similar to the proof of Proposition 1 in the lasso lecture notes. Assume that $X\tilde{\beta}$ is a nonnegative combination of exactly $n+1$ of $\pm X_j$, $j = 1, \ldots, d$. Then one of these $n+1$ vectors, denote it by $s_i X_i$ (where $s_i = \text{sign}(\tilde{\beta}_i)$) can be written as a linear combination of the others. Take an inner product with the lasso residual and use the subgradient optimality condition for the lasso to prove that the coefficients in this linear combination must sum to 1, and therefore, $s_i X_i$ is actually an affine combination of the others. Notice that this shows the affine span of the $n+1$ vectors in question is $(n-1)$-dimensional.

(d) A refinement of Carathéodory's is as follows: given a set $C \subseteq \mathbb{R}^k$, every element in its convex hull $\text{conv}(C)$ can be represented as a convex combination of $r+1$ elements of C, where r is the affine dimension of $\text{conv}(C)$.

Use this theorem and part (c) to prove that there exists a lasso solution $\check{\beta}$ with at most n nonzero coefficients. [2 pts]

2 Variance of least squares in nonlinear feature models [15 points]

In this exercise, we will examine the variance of least squares (in the underparametrized regime) and min-norm least squares (in the overparametrized regime) in nonlinear feature models. Recall for a response vector $Y \in \mathbb{R}^n$ and feature matrix $X \in \mathbb{R}^{n \times d}$, the min-norm least squares estimator $\hat{\beta} = (X^\mathsf{T} X/n)^+ X^\mathsf{T} Y/n$ has a variance component of its out-of-sample prediction risk (conditional on X) given by:

$$V_X(\hat{\beta}) = \frac{\sigma^2}{n} \mathrm{tr}(\hat{\Sigma}^+ \Sigma). \tag{1}$$

Here $\hat{\Sigma} = X^\mathsf{T} X/n$, and $\Sigma = \mathrm{Cov}(x_i)$, for an arbitrary row x_i of X (the rows all have the same distribution). Also, $\sigma^2 = \mathrm{Var}[y_i|x_i]$ is the noise variance. In lecture, we studied a linear feature model of the form

$$X = Z\Sigma^{1/2}, \tag{2}$$

for a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ and a random matrix $Z \in \mathbb{R}^{n \times d}$ that has i.i.d. entries with mean zero and unit variance. When $\Sigma = I$, which we will assume throughout this homework problem, recall that we proved that the variance (1) satisfies, under standard random matrix theory conditions, as $n, d \to \infty$ and $d/n \to \gamma \in (0, \infty)$,

$$V_X(\hat{\beta}) \xrightarrow{\text{as}} \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1 \\ \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{3}$$

(The result for $\gamma < 1$ actually holds regardless of Σ.) Instead, we can consider a nonlinear feature model of the form

$$X = \varphi(Z\Gamma^{1/2} W^\mathsf{T}), \tag{4}$$

for a covariance matrix $\Gamma \in \mathbb{R}^{k \times k}$, and a random matrix $Z \in \mathbb{R}^{n \times k}$ as before (except with k in place of d). Moreover, now $W \in \mathbb{R}^{d \times k}$ is a matrix of i.i.d. $N(0, 1/k)$ entries, and $\varphi : \mathbb{R} \to \mathbb{R}$ is a nonlinear function—called the activation function in a neural network context—that we interpret to act elementwise on its input.

There turns to be an uncanny connection between the asymptotic variance in linear and nonlinear feature models, which will you uncover via simulation in this homework problem. Attach your code as an appendix to this homework.

****Note: for parts (a)–(d) below, just consider isotropic features, so that $\Sigma = I$ in (1) and (2), and $\Gamma = I$ in (4).****

(a) Fix $n = 200$, and let $d = [\gamma n]$ over a wide range of values for γ (make sure your range covers both $\gamma < 1$ and $\gamma > 1$). each n, d, draw X from the linear feature model (2) and your choice of distribution for the entries of Z. Compute the finite-sample variance (1), and plot it, as a function of γ, on top of the asymptotic variance curve (3). To get a general idea of what this should look like, refer back to Figure 2 in the overparametrization lecture notes. [3 pts]

(b) For the same values of n, d, and $k = 100$, draw X from the nonlinear feature model (4), for three different choices of φ:

i. $\varphi(x) = a_1 \tanh(x)$;

ii. $\varphi(x) = a_2(x_+ - b_2)$;

iii. $\varphi(x) = a_3(|x| - b_3)$.

Here a_1, a_2, b_2, a_3, b_3 are constants that you must choose to meet the standardization conditions $\mathbb{E}[\varphi(G)] = 0$ and $\mathbb{E}[\varphi(G)^2] = 1$, for $G \sim N(0, 1)$. Produce a plot just as in part (a), with the finite-sample variances for choice of each activation function plotted in a different color, on top of the asymptotic variance curve (3) for the linear model case. Comment on what you find: do the nonlinear finite-sample variances lie close to the asymptotic variance for the linear model case? [9 pts]

(c) Now use a linear activation function $\phi(x) = ax - b$, and create a plot as in part (b) with the same settings (same values of n, d, k, and so on). What behavior do the finite-sample variances have as a function of γ? Is this surprising to you? Explain why what you are seeing is happening. [3 pts]

(d) As a bonus, in light of part (c), elaborate on why the results in part (b) are remarkable.

(e) As another (large) bonus, rerun the analysis in this entire problem but with a non-isotropic covariance Σ in (2), and Γ in (4). Extra bonus points if you properly recompute the asymptotic variance curves.

3 Implicit regularization buffet: choose your dish [15 points]

In this exercise, you get to choose between multiple options for studying implicit regularization. To be clear, each part (a), (b), and (c) below are worth equal points, and you only have to choose one (bonus points for doing more than one).

(a) Given any response vector $Y \in \mathbb{R}^n$ and feature matrix $X \in \mathbb{R}^{n \times d}$, let $\hat{\beta}^{\mathrm{rg}}(\lambda)$ denote the ridge estimator, which solves

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

for a tuning parameter $\lambda \geq 0$, and $\hat{\beta}_t^{\mathrm{gf}}$ denote the gradient flow estimator, which solves

$$\dot{\beta}(t) = X^\mathsf{T}(Y - X\beta(t)),$$

over time $t \geq 0$, subject to the initial condition $\beta(0) = 0$. Prove that these two satisfy the relationship, for any $\lambda \geq 0$:

$$\hat{\beta}^{\mathrm{rg}}(\lambda) = \mathbb{E}[\hat{\beta}^{\mathrm{gf}}(T_\lambda)], \quad T_\lambda \sim \mathrm{Exp}(1/\lambda),$$

where $\mathrm{Exp}(\mu)$ denotes the exponential distribution with mean μ. In other words, if we stop gradient flow at a random time that fluctuates around $1/\lambda$, then on average it will look like ridge with tuning parameter λ.

Hint: there are two ways to prove this. Either way starts by having you make the observation that for a function g, its Laplace transform evaluated at a λ, denoted $\mathcal{L}\{g\}(\lambda)$, can be written in terms of an expectation of g with respect to the $\mathrm{Exp}(1/\lambda)$ distribution. Then you have two paths you can pursue. The first path is to write down the explicit forms for the ridge and gradient flow solutions, and relate them using Laplace transforms. The second path is to start with the differential equation that defines gradient flow and take Laplace transforms, then relate this to the optimality conditions for ridge. Either way you will have to read up a little bit on Laplace transforms if you don't know much about them already.

(b) We saw in the ridge regression lecture that we can view the ridge estimator in terms of the min-norm least squares estimator on an augmented feature set, where we append a growing number of random features (independent of the response and given features) with suitably shrinking variance. Conduct a comprehensive simulation suite to (i) validate that this fact appears to be true for a very large number of auxiliary features, and (ii) investigate the speed of convergence: what happens as you vary the number of auxiliary features, down to a moderate or even small number?

Then, extend your simulations study to examine what happens when you use an augmented feature set (by appending random features) in other interpolating estimators, like the minimum ℓ_1 norm least squares estimator, or a neural network. Describe what you find. Does anything interesting come out?

(c) Conduct a literature search to identify an interesting example of implicit regularization in statistics or machine learning that is new—new just meaning that we did not cover it in class. This could be an example that is studied theoretically, or one that arises from methods or algorithms that are popular in practice. Describe in reasonable detail (1-2 pages) what is known about this example of implicit regularization.