Homework 3

Advanced Topics in Statistical Learning, Spring 2024

Due Friday March 22

1 Carathéodory's view on sparsity of lasso solutions [18 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 [4 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. [4 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 [6 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. [4 pts]

2 Variance of least squares in nonlinear feature models [30 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:

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

Here $\hat{\Sigma} = X^\mathsf{T} X/n$, and $\Sigma = \operatorname{Cov}(x_i)$, for an arbitrary row x_i of X (the rows all have the same distribution). Also, $\sigma^2 = \operatorname{Var}[y_i|x_i]$ 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.

For parts (a)–(d) below, 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. [6 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? [18 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. [6 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 The implicit regularization of gradient flow [25 points]

We will study gradient flow, as a continuous-time limit of the gradient descent path, in least squares regression. To build up motivation, consider gradient descent applied to the least squares regression problem

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

for a response vector $Y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{n\times d}$. For a given fixed step size $\epsilon > 0$, and for an initialization $\beta^{(0)} = 0$, gradient descent repeats the iterations:

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \frac{X^\mathsf{T}}{n}(Y - X\beta^{(k-1)}),$$

for $k = 1, 2, 3, \ldots$. Rearranging gives

$$\frac{\beta^{(k)} - \beta^{(k-1)}}{\epsilon} = \frac{X^\mathsf{T}}{n}(Y - X\beta^{(k-1)}),$$

and letting $k \to \infty$ and $\epsilon \to 0$, in such a way that $k\epsilon = t$, we get a continuous-time ordinary differential equation

$$\dot{\beta}(t) = \frac{X^\mathsf{T}}{n}(Y - X\beta(t)), \tag{5}$$

over time $t > 0$, subject to an initial condition $\beta(0) = 0$. We refer to the solution as the gradient flow path for least squares regression.

(a) Prove that the gradient glow path, the solution in (5), is [4 pts]

$$\hat{\beta}^{\text{gf}}(t) = (X^\mathsf{T}X)^+(I - \exp(-tX^\mathsf{T}X/n))X^\mathsf{T}Y.$$

Here A^+ is the Moore-Penrose generalized inverse of a matrix A, and $\exp(A) = I + A + A^2/2! + A^3/3! + \cdots$ is the matrix exponential of A. Note: you may use whatever properies of the matrix exponential that you want, as long as you clearly state them. (Also, you do not have to prove uniqueness of the solution in (5), you just have to plug in the above expression and show that it solves (5).)

(b) Let $X = \sqrt{n}US^{1/2}V^\mathsf{T}$ be a singular value decomposition, so that $X^\mathsf{T}X/n = VSV^\mathsf{T}$ is an eigendecomposition. Letting u_i, $i = 1, \ldots, p$ denote the columns of U, and s_i, $i = 1, \ldots, p$ the diagonal entries of S, prove that the vector of in-sample predictions from gradient flow are [4 pts]

$$X\hat{\beta}^{\text{gf}}(t) = \sum_{i=1}^p (1 - \exp(-ts_i))u_iu_i^\mathsf{T}Y.$$

(c) Recall, the vector of in-sample predictions from ridge regression with tuning parameter $\lambda > 0$,

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

can be written as

$$X\hat{\beta}^{\text{ridge}}(\lambda) = \sum_{i=1}^p \frac{s_i}{s_i + \lambda}u_iu_i^\mathsf{T}Y.$$

Thus both gradient flow and ridge perform a shrunken regression, by shrinking the eigenvalues of the empirical covariance matrix, but they use different underlying shrinkage maps, respectively:

$$g^{\text{gf}}(s,t) = 1 - \exp(-ts),$$
$$g^{\text{ridge}}(s,\lambda) = \frac{s}{s+\lambda}.$$

Plot g^{ridge} as a heatmap with s on the x-axis, and λ on the y-axis. Then, using the parametrization $t = 1/\lambda$, plot g^{gf} as a heatmap, again over s (x-axis) and λ (y-axis). Do you notice a similarity? [4 pts]

(d) Under the model

$$Y = X\beta_0 + \epsilon,$$
$$\text{where } \mathbb{E}[\epsilon] = 0,\ \text{Cov}(\epsilon) = \sigma^2 I,$$

and with X treated as fixed, prove that the estimation risk of gradient flow is [8 pts]

$$\mathbb{E}\|\hat{\beta}^{\text{gf}}(t) - \beta_0\|_2^2 = \sum_{i=1}^{p} \left(|v_i^\top \beta_0|^2 \exp(-2ts_i) + \frac{\sigma^2}{n} \frac{(1-\exp(-ts_i))^2}{s_i} \right),$$

where recall $X^\top X/n = V^\top S V$ is an eigendecomposition, and we use s_i, $i = 1, \ldots, p$ for the diagonal entries of S, and v_i, $i = 1, \ldots, p$ for the columns of V.

(e) Prove that, under the model from part (d), [5 pts]

$$\mathbb{E}\|\hat{\beta}^{\text{gf}}(1/\lambda) - \beta_0\|_2^2 \leq 1.6862 \cdot \mathbb{E}\|\hat{\beta}^{\text{ridge}}(\lambda) - \beta_0\|_2^2,$$

for any $\lambda > 0$. In words, the estimation risk of gradient flow at time $t = 1/\lambda$ is no more than 1.6862 times that of ridge at regularization parameter λ.

Hint: you may use the fact that the estimation risk of ridge is

$$\mathbb{E}\|\hat{\beta}^{\text{ridge}}(\lambda) - \beta_0\|_2^2 = \sum_{i=1}^{p} \left(|v_i^\top \beta_0|^2 \frac{\lambda^2}{(s_i+\lambda)^2} + \frac{\sigma^2}{n} \frac{s_i}{(s_i+\lambda)^2} \right).$$

Bonus: derive this. Another hint: you may use the fact that for all $x \geq 0$, it holds that

(i) $e^{-x} \leq 1/(1+x)$, and

(ii) $1 - e^{-x} \leq 1.2985 \cdot x/(1+x)$.

Bonus: prove these facts.