Math 542: Analysis of Variance and Regression
Final exam (take-home)

① Ridge regression.

Consider the setup of ℓ_2-regularized linear regression (a.k.a. *Tikhonov's* or *ridge* regression) discussed in the class. More precisely, the design vectors $\vec{x}_1, ..., \vec{x}_n \in \mathbb{R}^d$ are *fixed* and, as before,[1]

$$y_i = \langle \vec{x}_i, \theta^* \rangle + \xi_i, \quad i \in [n],$$

where $\sigma > 0$ is known, $\xi_i \sim \mathcal{N}(0,1)$ are i.i.d. noise realizations, and $\theta^* \in \mathbb{R}^d$ is unknown and to be estimated. As previously, we can rewrite the above identity in a compact matrix-vector form as

$$Y = \boldsymbol{X}\theta^* + \xi \tag{1}$$

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \vec{x}_1^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

Now, let $\|\cdot\|$ be the usual ℓ_2-norm (the square root of the sum of the squared entries of a vector). Define the empirical risk

$$L_n(\theta) := \frac{1}{n} \sum_{i \in [n]} (y_i - \langle \vec{x}_i, \theta \rangle)^2 = \frac{1}{n} \|Y - \boldsymbol{X}\theta\|^2$$

and the population risk (with expectation taken only over ξ_i's since $\vec{x}_i$'s are deterministic here):

$$L(\theta) := \mathbb{E}_\xi L_n(\theta) = \frac{1}{n} \|\boldsymbol{X}(\theta - \theta^*)\|^2 + \frac{d}{n}.$$

Note that θ^* is a minimizer of $L(\cdot)$, and for any θ, the *excess* population risk is a quadratic form[2]

$$L(\theta) - L(\theta^*) = \frac{1}{n} \|\boldsymbol{X}(\theta - \theta^*)\|^2 = \|\theta - \theta^*\|_{\boldsymbol{\Sigma}}^2$$

with matrix $\boldsymbol{\Sigma} := \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}$. Generally speaking, $\boldsymbol{\Sigma}$ does not have to be full-rank, and so the associated to it "prediction norm" $\|\cdot\|_{\boldsymbol{\Sigma}}$ might only be a seminorm, i.e. vanish for *nonzero* vectors; in particular, this is surely the case whenever $n < d$. In this problem, we do *not* assume that $n \geqslant d$.

- *We are free to just ignore the constant term $\frac{d}{n}$ in the population risk. Can you explain why?*

[1] For simplicity, we assume that $\sigma = 1$ here, i.e. the noise is "standardized."

[2] We write $\boldsymbol{\Sigma}$, rather than $\widehat{\boldsymbol{\Sigma}}_n$, for simplicity. We can get away with this since the design is deterministic anyway.

Recall the ridge estimate considered in the class:[3]

$$\widehat{\theta}_n^\lambda := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta) + \lambda \|\theta\|^2 \tag{2}$$

(Note that there is indeed a unique solution to this problem—why?) We shall bound its excess risk (working out some previously omitted details), then analyze a special regime of eigenvalue decrease.

1.1. Explicit form. Express $\widehat{\theta}_n^\lambda$ explicitly as a function of Y. *Hint: in the unregularized case* ($\lambda = 0$) *with* $\mathbf{\Sigma} \succ 0$, *the estimate used* $\mathbf{\Sigma}^{-1}$ *which might not exist now—but* $(\mathbf{\Sigma} + \lambda \boldsymbol{I})^{-1}$ *still does.*

1.2. Unbiasedness. Consider the regularized population risk minimizer:

$$\theta^\lambda := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} L(\theta) + \lambda \|\theta\|^2.$$

Derive θ^λ in explicit form, and show that $\widehat{\theta}_n^\lambda$ is its unbiased estimate (a special fact for linear models).

1.3. Variance term. Show that

$$\mathbb{E}\big[L(\widehat{\theta}_n^\lambda)\big] - L(\theta^\lambda) \leqslant \frac{d_\lambda(\mathbf{\Sigma})}{n}$$

where $d_\lambda(\mathbf{\Sigma}) := d_\lambda(\mathbf{\Sigma}) := \operatorname{tr}(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1})$ is called *the number of degrees of freedom (at level* λ*); here*

$$\mathbf{\Sigma}_\lambda := \mathbf{\Sigma} + \lambda \boldsymbol{I}.$$

Hint: use that $\operatorname{tr}(\boldsymbol{Q}^2) \leqslant \operatorname{tr}(\boldsymbol{Q})\, \lambda_{\max}(\boldsymbol{Q})$ *for any* $\boldsymbol{Q} \succeq 0$, *but be ready to explain how to prove this.*

1.4. Bias term, risk decomposition. Show that

$$L(\theta^\lambda) - L(\theta^*) \leqslant \lambda \|\theta^*\|^2 \tag{3}$$

Combine this result with the previous one to bound the excess risk as follows:

$$\mathbb{E}\big[L(\widehat{\theta}_n^\lambda)\big] - L(\theta^*) \leqslant \frac{d_\lambda(\mathbf{\Sigma})}{n} + \lambda \|\theta^*\|^2. \tag{4}$$

[3] We use a superscript to avoid possible confusion with a double subscript.

*② **Bias refinements in ridge regression.**

2.1. Refinement for small λ. In fact, the bias bound (3) is rather crude when λ is small. Identify the source of this the looseness and show the following improved bound:

$$L(\theta^\lambda) - L(\theta^*) \leqslant \lambda(\|\theta^*\|^2 - \|\theta^\lambda\|^2)$$
$$= \lambda \|\theta^*\|^2_{\boldsymbol{I}-\boldsymbol{J}_\lambda^2} \quad \text{where} \quad \boldsymbol{J}_\lambda := \boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}.$$

Simplify the last bound, by slightly roughening it, to

$$L(\theta^\lambda) - L(\theta^*) \leqslant 2\lambda^2 \|\theta^*\|^2_{\boldsymbol{\Sigma}_\lambda^{-1}}.$$

Explain why this last bound is always at least as strong as $2\lambda\|\theta^*\|^2$, i.e. twice the bound in (3).

Hint: note that $\boldsymbol{\Sigma}_\lambda$ commutes with $\boldsymbol{\Sigma}$, so we can express all related traces and matrix norms explicitly in terms of λ and the eigenvalues $\lambda_1, ..., \lambda_d$ of $\boldsymbol{\Sigma}$. E.g. for the degrees of freedom parameter:

$$d_\lambda(\boldsymbol{\Sigma}) = \sum_{k=1}^{d} \frac{\lambda_k}{\lambda_k + \lambda}.$$

2.2. Refinement for large λ. Note that as $\lambda \to \infty$, the first term in the right-hand side of (4) vanishes, but the bias term diverges. Clearly, this does not reflect what happens in reality: from (2) we see directly that $\theta^\lambda \to 0$ and $\widehat{\theta}_n^\lambda \to 0$ almost surely as $\lambda \to \infty$, and both the the excess risk and the bias converge to $L(0) - L(\theta^*) = \|\theta^*\|^2_{\boldsymbol{\Sigma}}$. Show the following bound (valid for any $\lambda \in [0, \infty]$):

$$L(\theta^\lambda) - L(\theta^*) \leqslant \lambda^2 \|\theta^*\|^2_{\boldsymbol{J}_\lambda \boldsymbol{\Sigma}_\lambda^{-1}}.$$

Observe that this bound is stronger than the one in **2.1**, and I do not mean the factor of 2 here.

*③ **Ridge regression in a nonparametric regime.**

In the setup of Problem **1**, consider the bound (4) from **1.4**. Assume that d is very large (or even infinite, if you prefer), and the eigenvalues $\lambda_1, \lambda_2, ...$ of $\mathbf{\Sigma}$ decrease, for a given $\alpha \geqslant 1$, as

$$\lambda_k = k^{-2\alpha}.$$

Let also $\|\theta^*\| \leqslant r$. Under these assumptions, show that the nearly best choice of λ for given α, r, n is

$$\lambda^* = c_{\alpha,r} n^{-\frac{2\alpha}{2\alpha+1}},$$

which results in $d_{\lambda^*} = asd$ the resulting excess risk bound is

$$\mathbb{E}[L(\widehat{\theta}_n^{\lambda^*})] - L(\theta^*) \leqslant C_{\alpha,r} n^{-\frac{2\alpha}{2\alpha+1}},$$

where $c_{\alpha,r}$ and $C_{\alpha,r}$ depend only on α and r, but not on n.

> *Hint: split the series*
>
> $$d_\lambda(\mathbf{\Sigma}) = \sum_{k=1}^{\infty} \frac{k^{-2\alpha}}{k^{-2\alpha} + \lambda}$$
>
> *into two parts: the "bulk" with the terms of nearly the same magnitude, and the "tail" where they rapidly decrease. Estimate the "tail" by replacing summation with integration.*

Discussion. This $n^{-\frac{2\alpha}{2\alpha+1}}$ convergence rate is, in fact, a common phenomenon in nonparametric functional regression;[4] two great texts on the topic are [Tsy09] and [Joh15] (available online). The larger is α, the smaller is the corresponding d_{λ^*}—the "effective dimension" of the parameter. In particular, $\alpha \to \infty$ corresponds to $d_{\lambda^*} = O(1)$ and the parametric $O(1/n)$ excess risk.[5] On the other hand, in the limit $\alpha \to 0$ we get no restriction of eigenvalues, and the bound becomes trivial.[6]

[4]Recall from the class that $k^{-2\alpha}$ is the rate of decrease for the Fourier coefficients of an α-differentiable function.

[5]As it turns out, when $\alpha \to \infty$ the bound does not depend on r as $\lim_{\alpha\to\infty} C_{\alpha,r} \equiv C$ for some *numerical* constant C.

[6]The assumption $\alpha \geqslant 1$ is technical; in fact, one may show that the results extend to $\alpha \geqslant 0$.

④ **Polynomial regression.** Linear regression can describe *seemingly* nonlinear dependencies. E.g., consider n noisy samples of unknown polynomial $\mathrm{p}(t)$ of degree $\leqslant d-1$ at $t_1 \neq ... \neq t_n \in [0,1]$:

$$\mathrm{y}(t_i) = \underbrace{\sum_{j \in [d]} \theta_j^* \varphi_j(t_i)}_{\mathrm{p}(t_i)} + \xi_i, \quad i \in [n], \tag{5}$$

where $\varphi_j(t) = t^{j-1}$, and $\theta_j^* \in \mathbb{R}^d$ is the corresponding coefficient in p. Clearly, this is (1) with

$$Y = \begin{bmatrix} \mathrm{y}(t_1) \\ \vdots \\ \mathrm{y}(t_n) \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \dots & \varphi_d(t_1) \\ \vdots & \vdots & & \vdots \\ \varphi_1(t_n) & \varphi_2(t_n) & \dots & \varphi_d(t_n) \end{bmatrix} = \mathbf{V}_{n,d}(t_1, ..., t_n)$$

where $\mathbf{V}_{n,d}$ *the rectangular Vandermonde matrix:*

$$\mathbf{V}_{n,d}(t_1, ..., t_n) := \begin{bmatrix} 1 & t_1 & \dots & t_1^{d-1} \\ \vdots & \vdots & & \vdots \\ 1 & t_n & \dots & t_n^{d-1} \end{bmatrix}.$$

Also, $\mathbf{V}_n(t_1, ..., t_n) := \mathbf{V}_{n,n}(t_1, ..., t_n)$ is known as *the square Vandermonde matrix* (of order n).

4.1. Nondegeneracy. Show that $\mathrm{rank}(\mathbf{V}_{n,d}(t_1, ..., t_n)) = d$ whenever $n \geqslant d$ and $t_1 \neq ... \neq t_n$.

Hint: I'm aware of two ways to solve this problem. One way is to first observe that it suffices to consider the square case $n = d$ (why?), and then prove the explicit formula

$$\det(\mathbf{V}_n(t_1, ..., t_n)) = \prod_{1 \leqslant i < j \leqslant n} (t_i - t_j),$$

whereby it follows that $\mathbf{V}_n(t_1, ..., t_n)$ is nonsingular if $t_1 \neq ... \neq t_n$ (and only in this case). The other way is to obtain a contradiction with the fundamental theorem of algebra (Gauss, 1799) in the form: "Any polynomial of degree d has $\leqslant d$ distinct complex roots."

4.2. Hilbert's matrix. Let $\mathbf{\Sigma}_n := \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}$ with $\boldsymbol{X} = \mathbf{V}_{n,d}(t_1, ..., t_n)$ as before, but now with

$$t_i = \frac{i}{n}, \quad i \in [n]. \tag{6}$$

Show that $\lim_{n \to \infty} \mathbf{\Sigma}_n = \mathbf{H}_d$ entrywise, where $\mathbf{H}_d$ is a matrix with entries $[\mathbf{H}_d]_{jk} = \frac{1}{j+k-1}$, that is

$$\mathbf{H}_d = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & & & \frac{1}{d} \\ \frac{1}{2} & \frac{1}{3} & & & \iddots & \\ \frac{1}{3} & & \ddots & \iddots & & \\ & & \iddots & & & \\ & \iddots & & & \ddots & \\ \frac{1}{d} & & & & & \frac{1}{2d-1} \end{bmatrix},$$

called the *Hilbert matrix* of order d. *Hint: don't forget the $\frac{1}{n}$ factor, which is also the grid step!*

4.3. Now, assume that instead of being fixed, $t_1, ..., t_n$ are sampled i.i.d. from $\mathsf{Uniform}([0,1])$. Argue that in this case, we are in the *random-design* linear regression setup, with $\mathbf{H}_d$ as the *population* covariance: $\mathbb{E}[\widehat{\mathbf{\Sigma}}_n] = \mathbf{H}_d$. (You don't need any more calculations on top of those in **4.2**.)

4.4. Let again $t_1, ..., t_n$ be on the regular grid with step $\frac{1}{n}$, cf. (6), and show that in this case,

$$[\mathbf{H}_d]_{jk} \leqslant [\mathbf{\Sigma}_n]_{jk} \leqslant [\mathbf{H}_d]_{jk} + \frac{1}{n}$$

in each entry. *Hint: play with the sum when appoximating it with an integral.*

*⑤ **Eigenvalue bounds.**

5.1. Absolute error. Show that

$$\|\mathbf{\Sigma}_n - \mathbf{H}_d\| \leqslant \frac{d}{n},$$

or: "all eigenvalues of $\mathbf{\Sigma}_n - \mathbf{H}_d$ are $\leqslant \frac{d}{n}$ in absolute value." To this end, use the following result:

Theorem 1 (Gershgorin circle theorem)**.** *For any eigenvalue $\lambda(A)$ of a complex $d \times d$ matrix A,*

$$\exists j \in [d]: \quad |\lambda(A) - A_{jj}| \leqslant \sum_{k \neq j} |A_{jk}|.$$

In words: "any eigenvalue must lie in at least one *Gershgorin's disc* centered at a diagonal entry of A, and with radius given by the sum of off-diagonal entries in the corresponding row (or column, since A and $A^\top$ have the same eigenvalues)."

Gershgorin's theorem is the most basic tool to estimate eigenvalues in terms of the matrix entries (which, generally, is a hard nonlinear problem), and oftentimes the only one available.

5.2. Eigenvalue estimates. Bound the eigenvalues of $\mathbf{H}_d$ as follows (they must be positive—why?):

$$\lambda_{\min}(\mathbf{H}_d) \lesssim \frac{\log(2d)}{d} \lesssim \lambda_{\max}(\mathbf{H}_d) \lesssim \log(2d).$$

Here $\lesssim$ hides a constant factor. (*Hint: trace is equal to the sum of eigenvalues.*) Observe that $\lambda_{\max}(\mathbf{H}_d) \geqslant 1$ (why?), and conclude that the condition number of $\mathbf{H}_d$ is $\gtrsim d/\log(2d)$.

5.3 Using the results of **5.1 — —5.2**, conclude that, neglecting the logarithmic factor, we need at least $n \gtrsim d^2$ to estimate $\lambda_{\min}(\mathbf{H}_d)$ by $\lambda_{\min}(\mathbf{\Sigma}_n)$ with a constant relative accuracy—say 10%—i.e. such that

$$|\lambda_{\min}(\mathbf{\Sigma}_n) - \lambda_{\min}(\mathbf{H}_d)| \leqslant 0.1\lambda_{\min}(\mathbf{H}_d).$$

Discussion. This is a very loose analysis: say, it is known that $\lambda_{min}(\mathbf{H}_d)$ is *exponentially small* in d; thus, in reality we need a way larger n (i.e., finer grid) to approximate $\mathbf{H}_d$ with a constant accuracy. However, our analysis already gives something worse than $n \asymp d$ expected from Bernstein's inequality, and demonstrates that *regular grid* is a bad choice when having to deal with polynomials.

References

[Joh15] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*, 2015.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.