

Final Examination

This is a 180 minute in-class final examination.

You may use two pieces of paper (double-sided), but otherwise this examination is closed book. While you may ask us questions if you find a question confusing, we've tried pretty hard to make the exam unambiguous and clear, so we're unlikely to say much.

Please respect the honor code.

All problems have equal weight. Some are (quite) straightforward. Others, not so much.

Some problems involve applications. But you do not need to know *anything* about the problem area to solve the problem; the problem statement contains everything you need.

The problems do *not* appear in order of increasing difficulty.

Name: _____

Stanford ID #: _____

F.1 Let A be an $m \times n$ matrix. A function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is a *norm* if it satisfies three conditions:

- i. f is nonnegative, i.e., $f(x) \geq 0$ for all x
- ii. $f(x) = 0$ if and only if $x = \mathbf{0}$ (definiteness)
- iii. f satisfies the triangle inequality, i.e., $f(x + y) \leq f(x) + f(y)$ for all n -vectors x, y .

Give necessary and sufficient conditions on A for the function

$$f(x) = \|Ax\|$$

to be a norm on $\mathbf{R}^n$.

F.2 In some prediction problems, given a feature mapping $\phi : \mathbf{R}^n \rightarrow \mathbf{R}^d$ and data points $(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}$, $i = 1, 2, \dots, m$, with $m < d$, we seek the smallest *interpolating predictor*, meaning the vector $w \in \mathbf{R}^d$ such that

$$y_i = \phi(x_i)^T w \tag{1}$$

for each $i = 1, \dots, m$ with the smallest norm $\|w\|$ among all vectors satisfying the interpolation condition (1). Formulate finding the smallest norm interpolant as a constrained least squares problem. Give one sufficient condition for your problem to have a solution.

F.3 (*Left and right inverses.*) For each of the following conditions, say that the statement *must* hold, meaning that it is always true; it *may* hold, meaning that sometimes it could be true but is not necessarily so; and that it *cannot* hold, meaning the statement is never true.

(a) A is a tall matrix with all non-zero entries. Then A has a left inverse.

(b) A is an $n \times 1$ matrix with at least one non-zero entry. Then A has a left inverse.

(c) A is a $1 \times n$ all zeros matrix. Then A has a right inverse.

(d) $A \in \mathbf{R}^{m \times n}$ is a tall matrix, $m > n$, with all non-zero entries. Then A has a right inverse.

(e) A is a wide matrix with independent rows. Then A has a right inverse.

F.4 Regularization. The *regularized* least squares problems solves

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda \|x\|^2 \quad (2)$$

for $\lambda > 0$. Let $\hat{x}_\lambda$ denote the minimizer.

(a) Give as explicit a formula as you can for $\hat{x}_\lambda$.

(b) Does the problem (2) always have a unique solution? Why or why not?

(c) Suppose that the matrix $A \in \mathbf{R}^{m \times n}$ and vector $b \in \mathbf{R}^m$ come from paired measurements $(a_i, b_i) \in \mathbf{R}^n \times \mathbf{R}$ of a system of interest, and we wish to choose the $\lambda > 0$ that yields the best predictor $\hat{x}_\lambda$ for future data from the system. You have a collection $\Lambda = \{\lambda_1, \dots, \lambda_k\}$ of potential values to choose for λ . Describe in at most a single sentence how you would (approximately) find the best $\lambda \in \Lambda$.

F.5 Polynomial smoothing. Consider fitting a piecewise degree d polynomial function $f : \mathbf{R} \rightarrow \mathbf{R}$ to observed data $(x_i, y_i) \in \mathbf{R} \times \mathbf{R}$, $i = 1, \dots, m$. We define the j th polynomial by

$$\hat{f}_j(x) = a_{0,j} + a_{1,j}x + \dots + a_{d,j}x^d$$

for $j = 1, 2, \dots, k$, where $a_j = (a_{0,j}, \dots, a_{d,j}) \in \mathbf{R}^{d+1}$ are the coefficients of the polynomial terms. Then for $k - 1$ knot locations $l_1 < l_2 < \dots < l_{k-1}$, we define the piecewise function

$$f(x) = \hat{f}_j(x) \text{ when } l_{j-1} < x \leq l_j, \quad (3)$$

for $j = 1, \dots, k$, where we use the abuse of notation that $l_0 = -\infty$ and $l_k = +\infty$. Consider fitting a piecewise polynomial of the form (3) so that $\hat{f}$ minimizes the sum of squared errors

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \dots + (y_m - f(x_m))^2$$

over functions agreeing to first order at each of the knot locations l_j , that is, satisfying the zeroth and first order equalities

$$\hat{f}_j(l_j) = \hat{f}_{j+1}(l_j) \text{ and } \hat{f}'_j(l_j) = \hat{f}'_{j+1}(l_j)$$

at each knot l_j , $j = 1, \dots, k - 1$. Assume without loss of generality that $x_1 < x_2 < \dots < x_m$.

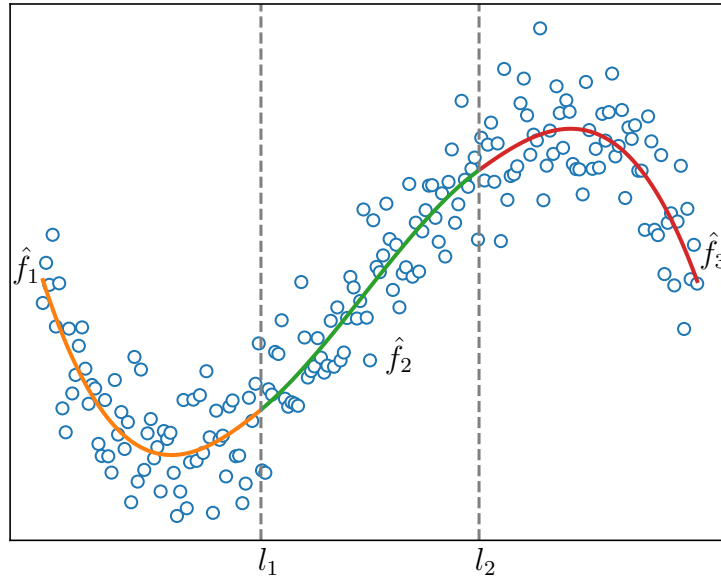


Figure 1: Example of piecewise polynomial fit with $k = 2$ knots at locations l_1 and l_2 .

(a) Explain how to formulate this as a constrained least squares problem of the form

$$\begin{aligned} & \text{minimize}_{\theta} \quad \|W\theta - y\|^2. \\ & \text{subject to} \quad C\theta = \mathbf{0} \end{aligned}$$

Describe the data matrix W , the constraint matrix C , the vector θ , and the targets y . Identify the dimensions of each component W, C, θ, y of your problem. It may be useful to use the vector functions $\phi : \mathbf{R} \rightarrow \mathbf{R}^{d+1}$

$$\phi(x) = (1, x, x^2, \dots, x^d) \text{ and } \phi'(x) = (0, 1, 2x, \dots, d \cdot x^{d-1}).$$

(b) Consider solving the KKT conditions

$$\begin{bmatrix} 2W^T W & C^T \\ C & \mathbf{0} \end{bmatrix} \begin{bmatrix} \theta \\ z \end{bmatrix} = \begin{bmatrix} 2W^T y \\ 0 \end{bmatrix}$$

for your problem. Give a few sentences about structure in W and/or C might you leverage to solve this more efficiently than naively inverting the KKT matrix? For example, (approximately) how many FLOPs does inverting $W^T W$ require?

F.6 Finding physical laws with regression. We can write many physical laws as polynomial equations. For example, Newton's second law is that $F = MA$ (force is mass times acceleration); the ideal gas law is that $PV = NRT$ (pressure times volume equals number of moles times ideal gas constant times temperature), kinetic energy is $K = \frac{1}{2}MV^2$, where M is mass and V velocity. Suppose we have k physical quantities, represented by k positive values $A_1 > 0, A_2 > 0, \dots, A_k > 0$, and we believe they are related by the equality

$$1 = c \cdot A_1^{\beta_1} \cdot A_2^{\beta_2} \cdots A_k^{\beta_k},$$

where $c > 0$ and $\beta_1, \dots, \beta_k \in \mathbf{R}$ are unknown scalars (but which we believe are non-zero).

We are given m experimental measurements of vectors a_i containing realizations of the A variables, i.e.,

$$a_i = (A_{1,i}, \dots, A_{k,i}) \in \mathbf{R}_{+}^k,$$

so the j th component of a_i is $A_{j,i} > 0$, for $i = 1, \dots, m$. How you might find the values $c, \beta_1, \dots, \beta_k$ by solving an ordinary or constrained least-squares problem involving the vectors $a_1, a_2, \dots, a_m$? Explain any transformations you make to the problem data and the size of your problem. *Hint.* Think about transformations to linearize the equation $1 = c \cdot A_1^{\beta_1} \cdots A_k^{\beta_k}$.

F.7 *Ranking sports teams.* Consider a sports league in which n teams play against one another; because of the size of the league, some teams never play one another, and no team plays another more than once. At the end of the season, we construct an $n \times n$ win/loss matrix $A \in \mathbf{R}^{n \times n}$ where

$$A_{ij} = \begin{cases} 1 & \text{if team } i \text{ defeated team } j \\ 0 & \text{if team } i \text{ lost to or never played team } j \\ 1/2 & \text{if teams } i \text{ and } j \text{ tied.} \end{cases}$$

Assume each team played a total of $m > 0$ games (no more, no fewer), and let $\mathbf{1} \in \mathbf{R}^n$ be the all-ones vector of size n .

(a) Give a one or two sentence interpretation of the vectors $A\mathbf{1}$ and $\frac{1}{m}A\mathbf{1}$.

(b) Now assume there were no ties (e.g., games go into overtime if tied, and teams play until one wins). Give a short (few sentence) interpretation of the vectors $A(\frac{1}{m}A\mathbf{1})$ and $(\frac{1}{m}A)^2\mathbf{1}$.