STATS 305 A - Lecture 3

John Duchi
Scribed by Michael Howes

09/28/21

Contents

1 Announcements

- Homowork 1 to be posted tonight. Due in two weeks. An email notice will be sent.
- TA's office hours times available on the website.
- John's office hours TBD.
- First Etude will probably be posted on Friday. Each etude will be checked/attempted by the TAs.
- R and python and maybe Julia will be supported for assignments.

2 Recap/Linear Algebra (SVD)

2.1 SVD in statistics

Recall that if $A \in \mathbb{R}^{m \times n}$ and $m \geq n$ (A is tall), then A has a singular value decompsition (SVD) as

$$A = U \Sigma V^T,$$

where $U \in \mathbb{R}^{m\times n}$ (same size as A and thus also tall), $\Sigma = \text{diag}(s_1, \ldots, s_n) \in \mathbb{R}^{n\times}$ and $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$, $V \in \mathbb{R}^{n\times n}$ and $U^TU = V^TV = I_n$.

We can give a statistical interpretation of the SVD. Suppose we have the linear model $Y = X\beta + \varepsilon$ where $\varepsilon \sim (0, \sigma^2 I_n)$, $X \in \mathbb{R}^{n\times d}$ and $n \geq d$ (X is tall and we have more observations than variables). Then $X = U\Sigma V^T$ with $\Sigma = \text{diag}(s_1, \ldots, s_d)$, $U = [u_1, \ldots, u_d]$ and $v = [v_1, \ldots, v_d]$. Our goal is to recover information about β from the data (X, Y).

The components of β in the v_d directions are "hard" to infer anything about. To make this precise, consider

$$\beta_0 \text{ and } \beta_t = \beta_0 - tv_d.$$

Then

$$\begin{aligned} X\beta_0 - X\beta_t &= -tXv_d \\ &= -tU \begin{bmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & s_d \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \\ &= -ts_d u_d, \end{aligned}$$

since $Vv_d = e_d$. Thus we need $t \sim \frac{1}{s_d}$ to induce substantial changes to the observed Y. Constrast this with making inferences in the v_1 direction. If $\beta_t = \beta_0 - tv_1$, then $X\beta_0 - X\beta_t = ts_1u_1$. Thus changes in the v_1 direction give us relatively big changes in Y.

Note that the singular values are always non-negative.

2.2 A concrete exmaple

[In answer to a question] Historically linear models were developed for agriculture where Y_i = yield in kg of field i and

$$X = \begin{bmatrix} — & X_1^T & — \\ & \vdots & \\ — & X_n^T & — \end{bmatrix},$$

where $X_i \in \{0, 1\}^d$ encodes the treatments given to field i and in particular $X_{i,j} = 1$ if treatment j was used in field i.

2.3 Geometric interpretation of SVD

[In answer to another question] Suppose $A = U\Sigma V^t$, then $Ax = U\Sigma Vx$, thus we can calculate Ax in three successive steps.

(Step 1) First

$$V^Tx = \begin{bmatrix} v_1^Tx \\ \vdots \\ v_n^Tx \end{bmatrix}$$

are the coordinates of x in the basis V.

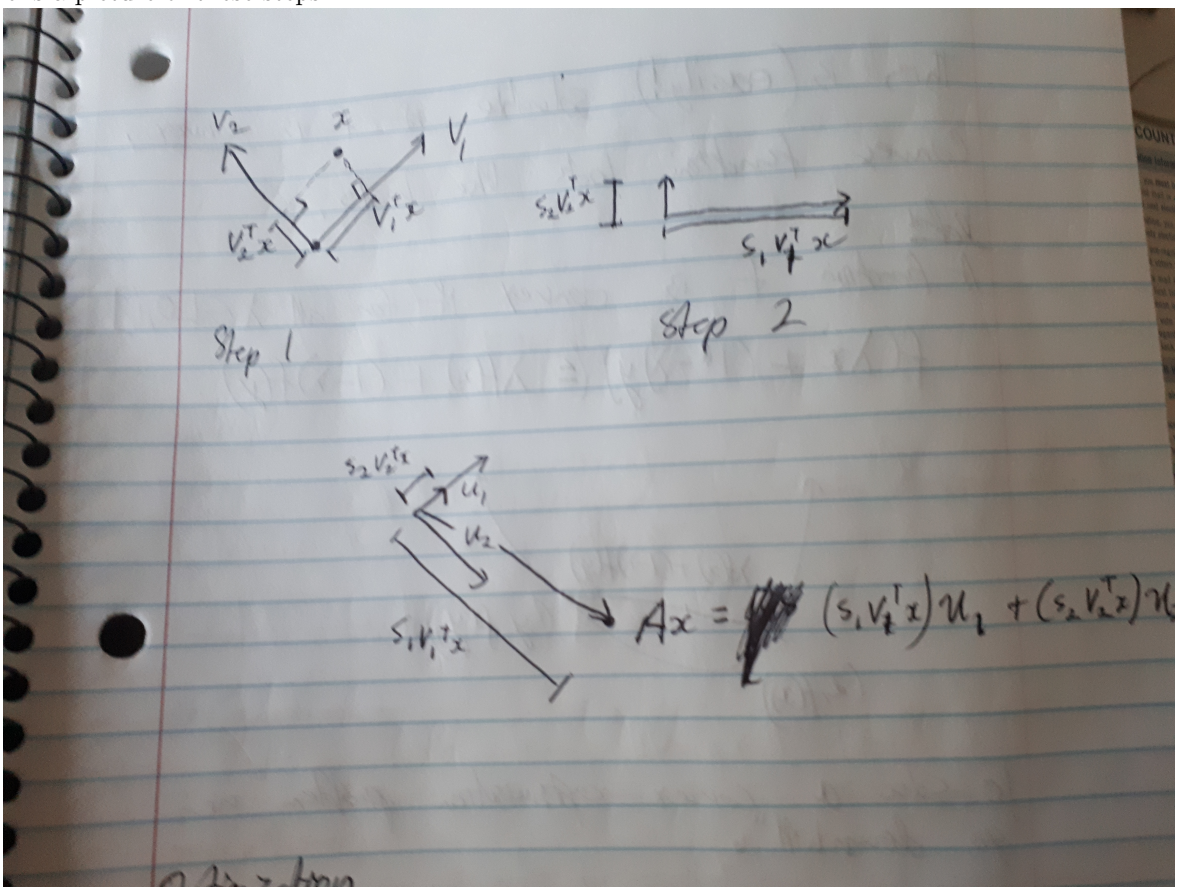
(Step 2) We apply s_i gains to the components of Vx

$$\Sigma V^Tx = \begin{bmatrix} s_1v_1^Tx \\ \vdots \\ s_nv_n^Tx \end{bmatrix}.$$

(Step 3) We use these values to define a new vector using the basis U.

$$Ax = U\Sigma V^T x = U \begin{bmatrix} s_1 v_1^T x \\ \vdots \\ s_n v_n^T x \end{bmatrix} = \sum_{i=1}^{n} (s_i v_i^T x) u_i.$$

Here is a picture of these steps.



3 Optimisation Basics

3.1 Convex Optimisation

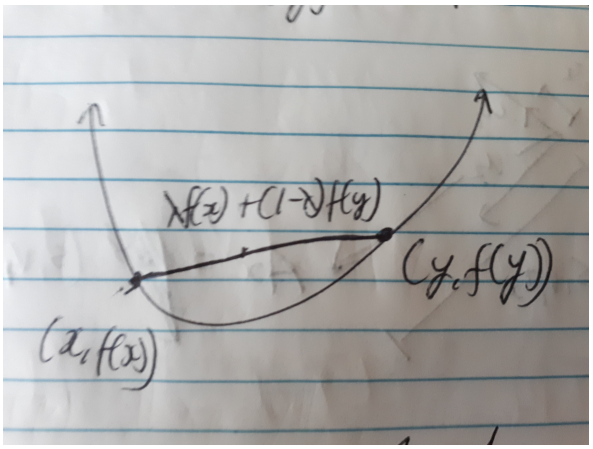
Optimisation is about problems of the form minimize $f(x)$ s.t $h(x) = 0$ and $g(x) \leq 0$. The function f is our objective, x is out variable, h is an equality constraint and g are inequality constraints. In this class we will mostly consider unconstrained cases.

This is solvable (easily!) is f is *convex*. If f isn't *convex* we are in trouble. Convex functions look like bowls.

Definition 1. A function f is convex if for all $\lambda \in [0, 1]$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

That is if we graph f, then the line between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f (bowl shaped). (See picture)

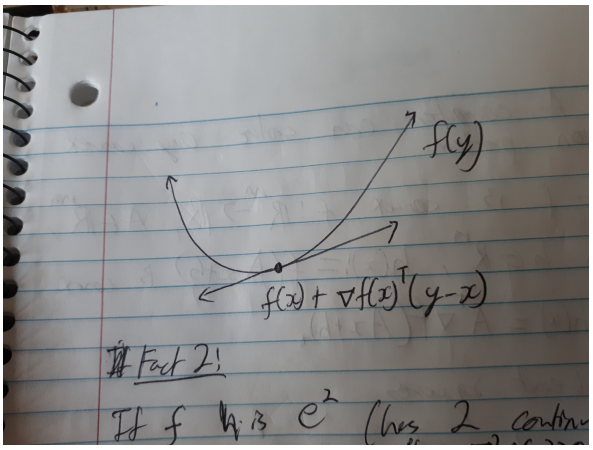


To solve a convex optimisation problem we "go downhill".

Fact If f is differentiable, then f is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x),$$

for all x and y. That is the line tangent to f at x lies below the graph of f. See picture



Fact 2 If f is $\mathcal{C}^2$ (has 2 continuous derivatives), then f is convex iff $\nabla^2 f(x) \succeq 0$ for all x where $\nabla^2 f(x) = \left[\frac{\partial^2}{\partial x_i \partial x_j} f(x)\right]_{i,j=1}^n$. Also for an $n \times n$ matrix $A \succeq 0$ means A is positive semi-definite i.e. $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$.

Proposition 1. *If f is convex, then x^* minimizes f if and only if $\nabla f(x^*) = 0$.*

Proof. If $\nabla f(x^*) = 0$, then for all y,

$$f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x) = f(x^*).$$

Thus $f(x^*)$ is the minimum value of f. The converse is similar. □

Fact 3 If $f : \mathbb{R}^m \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, then $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(x) := f(Ax+b)$ is convex and

$$\nabla h(x) = A^T \nabla f(Ax + b).$$

3.2 Least squares example

[An answer to a student's question] Suppose we are given the data $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$. We have the variable $\beta \in \mathbb{R}^d$ and we wish to minimize

$$L(\beta) = \frac{1}{2}\|X\beta - Y\|_2^2.$$

The function L is convex in β and $\nabla L(\beta) = X^T(X\beta - Y)$. Setting this equal to zero gives $X^TX\beta = X^TY$. These equations are called the normal equations. Thus the minimises of L are the solutions to the normal equations.

If X has rank d, then X^TX is invertible and

$$\beta = (X^TX)^{-1}X^TY,$$

is the unique minimizer of L.

3.3 Geometry of convex functions in higher dimensions

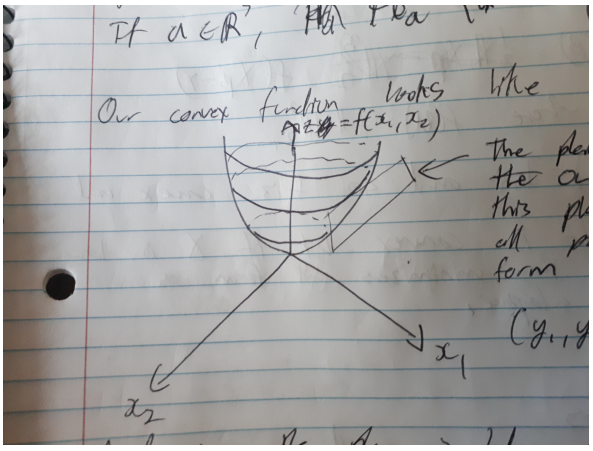
Let's revisit $f(y) \geq f(x) + \nabla f(x)^T(y-x)$ where $f : \mathbb{R}^2 \to \mathbb{R}$ and so we can graph the points $(y_1, y_2, f(y_1, y_2))$ in $\mathbb{R}^3$. The function f being convex means that this graph looks like a bowl. The set of all points

$$\left(y_1, y_1, f(x_1, x_2) + \nabla f(x_1, x_2)^T \begin{bmatrix} y_1 - x_1 \\ y_2 - x_2 \end{bmatrix}\right),$$

is the plane that is tangential to the graph of f at the point $(x_1, x_2, f(x_1, x_2))$. Since f is convex and looks like a bowl, we can think of putting a piece of paper that just touches a point on the underside of the bowl. By convexity this piece of paper never goes inside the bowl and thus

$$f(y) = f(y_1, y_2) \geq (x_1, x_2) + \nabla f(x_1, x_2)^T \begin{bmatrix} y_1 - x_1 \\ y_2 - x_2 \end{bmatrix} = f(x) + \nabla f(x)^T(y - x).$$

See also this picture



These geometric pictures are useful and important but not something we will be tested on. Note that convex functions may have more than one minimum.

3.4 More than one minimum

See handout on the course webpage for details. In our least squares example we could write

$$L(\beta) = \frac{1}{2}\|X\beta - Y\|_2^2,$$

as $L(\beta) = f(X\beta - Y)$ where $f(u) = \frac{1}{2}\|u\|_2^2$. Since $\nabla^2 f(u) = I_n$, f is convex and thus L is also convex. But if X is rank $r < d$, then there will be multiple solutions to the normal equations and hence multiple minimizers of L. These solutions will form an affine subspace. In general the minimizers of a convex function form a convex set.

3.5 Projections

Lets consider another problem when we can use the tools of convex optimisation. Given $a_1, \dots, a_k \in \mathbb{R}^d$ we wish to find the closest point to v in $\text{span}\{a_i\}_{i=1}^k$. Write $A = [a_1, \dots, a_n]$, then $x \in \text{span}\{a_i\}_{i=1}^n$ if and only if $x = A\lambda$ for some $\lambda \in \mathbb{R}^k$. Thus our problem is equivalent to the constrained optimisation

$$\min_{x,\lambda} \frac{1}{2}\|v - x\|_2^2 \text{ s.t. } x = A\lambda.$$

But we can rewrite this as an unconstrained optimisation problem

$$\min_{\lambda} \frac{1}{2}\|v - A\lambda\|_2^2.$$

We can calculate $\nabla_\lambda \frac{1}{2}\|v - A\lambda\|_2^2 = A^T(A\lambda - v)$. If A is full rank, then setting ∇_λ equal to 0 we get $\lambda = (A^TA)^{-1}A^Tv$ and thus $x = A(A^TA)^{-1}A^Tv$. We can also solve this problem by using the SVD of A.

If $A = U\Sigma V^T$ is the SVD of A, then

$$\begin{aligned}
\Pi_A &:= A(A^TA)^{-1}A^T \\
&= U\Sigma V^T(V\Sigma U^T U\Sigma V^T)^{-1}V\Sigma U^T \\
&= U\Sigma V^T(V\Sigma^2 V^T)^{-2}V\Sigma U^T \\
&= U\Sigma V^T V\Sigma^{-2}V^T V\Sigma U^T \\
&= UU^T.
\end{aligned}$$

We can also see this because $\text{span}(A) = \text{span}(U)$ and projection onto $\text{span}\{u_1, \dots, u_k\}$ is

$$\sum_{i=1}^k u_i(u_i^T x) = UU^T x.$$

3.6 Summary

This is all the optimisation we will need in this class. It is okay if it was unfamiliar. The upshot is that if we can frame a problem as a convex optimisation problem, then the computer can solve it.

4 Review of distributions

A random variable/vector $X \in \mathbb{R}^d$ with a density f or a probability mass function (p.m.f) p has expectation/mean

$$\mathbb{E}[X] = \begin{cases} \sum xp(x) & \text{if } X \text{ has a p.m.f } p, \\ \int xf(x)dx & \text{if } X \text{ has a density } f. \end{cases}$$

The random variable X aso has a covariance matrix

$$\text{Cov}(X) = V(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] = [\text{cov}(X_i, X_j)]_{i,j=1}^d,$$

where $\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]$. Note that $V(X) \succeq 0$ since

$$u^T V(X) u = \mathbb{E}\left[(u^T(X - \mathbb{E}X))^2\right] \geq 0.$$

4.1 Normal Distributions

$X \sim \mathcal{N}(\mu, \Sigma)$ means that X is normally distributed with expectation μ and covariance Σ. This mean X has density

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

assuming that $\Sigma \succ 0$, that is Σ is positive definite.

Some facts about normal distributions. If $X \sim \mathcal{N}(0, I)$ and $Y = Ax + b$ then $Y \sim \mathcal{N}(b, AA^T)$. A "proof" $\mathbb{E}[Y] = b$ and

$$\text{Cov}(Y) = \mathbb{E}[(AX)(AX)^T] = A\mathbb{E}[XX^T]A^T = AA^T.$$

A consequence of this is that normals are rotationally invariant. That is if $U \in \mathbb{R}^{d\times d}$ is orthogonal ($UU^T = I_d$) and $Z \sim \mathcal{N}(0, I)$, then $UZ \sim \mathcal{N}(0, I)$. That is UZ and Z have the same distriution.

What does a normal distribution look like?

If $\Sigma = V\Lambda V^T$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$, then the density of $X \sim \mathcal{N}(0, \Sigma)$ is proportional to $\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$. We know that $\Sigma^{-1} = V\Lambda^{-1}V^T$. Thus the level sets of $f(x)$ are

$$\{x : x^T\Sigma^{-1}x = \text{ constant}\} = \{x : (Vx)^T\Lambda^{-1}(Vx) = \text{ constant}\}.$$

These sets are elipses with axis in the direction $v_1, \ldots, v_d$ and length $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}$.