ORIE 4741: Learning with Big Messy Data

Underdetermined Least Squares and Quadratic Regularization

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Announcements 10/5/2021

- section this week: generalization and validation
- hw3 due next week, Friday 10am
    - save slip days for emergencies
- project peer reviews due Sunday 11:59pm
- iClicker not working? alas, best bet is to buy the app...

Announcements 10/7/2021

- quiz opens at noon today (Thursday), closes noon Saturday; take it before your fall break begins!
- project peer reviews due Sunday 11:59pm
- hw3 due next week, Friday 10am
    - save slip days for emergencies
- section next week (W only): advanced scikit-learn

Poll: fall break

For fall break, I'm

A. traveling starting Thursday
B. traveling starting Friday
C. traveling starting Saturday
D. staying in Ithaca
E. other

Poll: project presentations

I'd prefer to do the project presentations

A. live
B. as a video recording

Linear algebra review

Definition

The **null space** of a matrix $X : \mathbf{R}^{n \times d}$ is

$$\mathbf{nullspace}(X) = \{w \in \mathbf{R}^d : Xw = 0\}$$

(The all-zero vector 0 is always in the null space.)

The following conditions are equivalent:

- $\mathbf{nullspace}(X) = \{0\}$
- If $Xw = 0$, then $w = 0$
- The columns of X are linearly independent
- $\forall z \in \mathbf{R}^n$, if $Xw = z$ and $Xw' = z$, then $w = w'$
- X has a left inverse

Notation: standard basis vectors

- e_1 is the first standard basis vector $(1, 0, \ldots, 0)$
- e_2 is the second standard basis vector $(0, 1, 0, \ldots, 0)$
- $\{e_1, \ldots, e_d\}$ form the standard basis in $\mathbf{R}^d$

What if the Gram matrix is not invertible?

- Least squares objective:

$$\text{minimize} \qquad \|y - Xw\|^2$$

- Normal equations:

$$X^T X w = X^T y$$

- Solution if $X^T X$ is invertible:

$$w = (X^T X)^{-1} X^T y$$

Poll: rank-deficient normal equations

Normal equations:

$$X^T X w \;\; = \;\; X^T y$$

Q: if $X^T X$ is not invertible, do the normal equations still define the solution?

A. yes

B. no

Poll: rank-deficient normal equations

Normal equations:

$$X^T X w \;=\; X^T y$$

Q: if $X^T X$ is not invertible, do the normal equations still define the solution?

A. yes

B. no

A: yes! we derived them with no assumptions.

Outline

The SVD

Non-uniqueness

Quadratic regularization

The Singular Value Decomposition (SVD)

suppose $d \leq n$. SVD rewrites $X \in \mathbf{R}^{n \times d}$ in terms of easier matrices:

- $X = U \Sigma V^T$
- $U \in \mathbf{R}^{n \times d}$ is orthogonal: $U^T U = I_d$
- $V \in \mathbf{R}^{d \times d}$ is orthogonal: $V^T V = V V^T = I_d$
- $\Sigma \in \mathbf{R}^{d \times d}$ is diagonal and nonnegative:
  - $\Sigma_{ii} \geq 0$ for $i = 1, \ldots, d$
  - $\Sigma_{ij} = 0$ for $i \neq j$

The Singular Value Decomposition (SVD)

suppose $d \leq n$. SVD rewrites $X \in \mathbf{R}^{n \times d}$ in terms of easier matrices:

- $X = U \Sigma V^T$
- $U \in \mathbf{R}^{n \times d}$ is orthogonal: $U^T U = I_d$
- $V \in \mathbf{R}^{d \times d}$ is orthogonal: $V^T V = V V^T = I_d$
- $\Sigma \in \mathbf{R}^{d \times d}$ is diagonal and nonnegative:
  - $\Sigma_{ii} \geq 0$ for $i = 1, \ldots, d$
  - $\Sigma_{ij} = 0$ for $i \neq j$

use the SVD (in python, `scipy.linalg.svd(X, full_matrices=False)`)

```
U,S,V = svd(X)
```

can compute *SVD* factorization of X in $\mathcal{O}(nd^2)$ flops

Thin SVD

to make **thin SVD**, delete zeros from Σ

- $r = \mathrm{Rank}(X)$
- $X = U \Sigma V^T$
- $U \in \mathbf{R}^{n \times r}$ has orthogonal columns: $U^T U = I_r$
- $V \in \mathbf{R}^{d \times r}$ has orthogonal columns: $V^T V = I_r$
- $\Sigma \in \mathbf{R}^{r \times r}$ is diagonal and positive:
  - $\Sigma_{ii} > 0$ for $i = 1, \ldots, r$
  - $\Sigma_{ij} = 0$ for $i \neq j$

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$X^T X w \;\; = \;\; X^T y$$

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$\begin{aligned} X^T X w &= X^T y \\ V\Sigma^2 V^T w &= V\Sigma U^T y \end{aligned}$$

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$\begin{aligned} X^T X w &= X^T y \\ V\Sigma^2 V^T w &= V\Sigma U^T y \\ \Sigma^{-2} V^T V\Sigma^2 V^T w &= \Sigma^{-2} V^T V\Sigma U^T y \end{aligned}$$

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$\begin{aligned} X^T X w &= X^T y \\ V\Sigma^2 V^T w &= V\Sigma U^T y \\ \Sigma^{-2} V^T V \Sigma^2 V^T w &= \Sigma^{-2} V^T V \Sigma U^T y \\ V^T w &= \Sigma^{-1} U^T y \end{aligned}$$

can't solve (V^T not invertible, solution not unique. . .)

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$\begin{aligned} X^T X w &= X^T y \\ V\Sigma^2 V^T w &= V\Sigma U^T y \\ \Sigma^{-2} V^T V \Sigma^2 V^T w &= \Sigma^{-2} V^T V \Sigma U^T y \\ V^T w &= \Sigma^{-1} U^T y \end{aligned}$$

can't solve (V^T not invertible, solution not unique. . .)
guess $w = V\Sigma^{-1} U^T y = \sum_{i=1}^{d} v_i \sigma_i^{-1} u_i^T y$:

SVD for least squares

if $X = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$ is the thin SVD, then

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

normal equations are

$$\begin{aligned} X^T X w &= X^T y \\ V\Sigma^2 V^T w &= V\Sigma U^T y \\ \Sigma^{-2} V^T V\Sigma^2 V^T w &= \Sigma^{-2} V^T V\Sigma U^T y \\ V^T w &= \Sigma^{-1} U^T y \end{aligned}$$

can't solve (V^T not invertible, solution not unique...)
guess $w = V\Sigma^{-1}U^T y = \sum_{i=1}^d v_i \sigma_i^{-1} u_i^T y$:

$$V^T w = V^T V\Sigma^{-1} U^T y = \Sigma^{-1} U^T y$$

so we've found a solution (without assuming invertibility)!

Demo: SVD

https://github.com/ORIE4741/demos/SVD.ipynb

Review: methods for least squares

	GD	SGM	Gram GD	Parallel GD	QR or SVD
initial	0	0	nd^2	nd^2/P	nd^2
per iter	nd	$\|S\|d$	d^2	d^2	0

(numbers in flops, omitting constants)

- gradient descent (most flexible, $O(nd)$ flops per iteration)
- QR factorization (most efficient exact solution method, $O(nd^2)$ flops)
- SVD factorization (exact solution method, works for underdetermined problems, $O(nd^2)$ flops)
- backslash command uses either QR or SVD to ensure stability + speed

Outline

The SVD

Non-uniqueness

Quadratic regularization

Poll: uniqueness

Normal equations:

$$X^T X w \;=\; X^T y$$

Q: is the solution to the normal equations always unique?

A. yes

B. no

Poll: uniqueness

Normal equations:

$$X^T X w \;=\; X^T y$$

Q: is the solution to the normal equations always unique?

A. yes

B. no

A: no, if $X^T X$ is not invertible, the solution is not unique!
if $\mathrm{Rank}(X^T X) < d$, then for some $v \neq 0$, $X^T X v = 0$.
so if $X^T X w = X^T y$, then $X^T X(w + \alpha v) = X^T y$ for any $\alpha \in \mathbf{R}$.

Poll: uniqueness

Normal equations:

$$X^T X w = X^T y$$

Q: is the solution to the normal equations always unique?

A. yes

B. no

A: no, if $X^T X$ is not invertible, the solution is not unique!
if $\mathrm{Rank}(X^T X) < d$, then for some $v \neq 0$, $X^T X v = 0$.
so if $X^T X w = X^T y$, then $X^T X(w + \alpha v) = X^T y$ for any $\alpha \in \mathbf{R}$.
Q: is non-uniqueness a problem for a predictive model?

A. yes

B. no

Example: non-uniqueness

- goal: predict cancer risk from mutations in genes
- X_{ij} is 1 if person i has a mutation in gene j
- genes 1 and 2 vary together: every person with a mutation in gene 1 has one in gene 2, too, and vice versa
- so the first and second column of X are identical: $X_{1:} = X_{2:}$

Example: non-uniqueness (II)

$$X_{1:} = X_{2:}$$

- suppose our least squares solution is w
- $w' = w + \alpha e_1 - \alpha e_2$, for $\alpha \in \mathbf{R}$, makes the same predictions:

$$\begin{aligned} Xw' &= X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\ &= Xw + \alpha(X_{1:} - X_{2:}) = Xw \end{aligned}$$

- now suppose a new person x arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).

Example: non-uniqueness (II)

$$X_{1:} = X_{2:}$$

- suppose our least squares solution is w
- $w' = w + \alpha e_1 - \alpha e_2$, for $\alpha \in \mathbf{R}$, makes the same predictions:

$$\begin{aligned} Xw' &= X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\ &= Xw + \alpha(X_{1:} - X_{2:}) = Xw \end{aligned}$$

- now suppose a new person x arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).

Q: do w and w' make the same prediction?

A. yes
B. no

Example: non-uniqueness (II)

$$X_{1:} = X_{2:}$$

- suppose our least squares solution is w
- $w' = w + \alpha e_1 - \alpha e_2$, for $\alpha \in \mathbf{R}$, makes the same predictions:

$$\begin{aligned} Xw' &= X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\ &= Xw + \alpha(X_{1:} - X_{2:}) = Xw \end{aligned}$$

- now suppose a new person x arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).

Q: do w and w' make the same prediction?

A. yes
B. no

Q: what criteria might you pick to choose a good w?

Example: non-uniqueness (II)

$$X_{1:} = X_{2:}$$

- suppose our least squares solution is w
- $w' = w + \alpha e_1 - \alpha e_2$, for $\alpha \in \mathbf{R}$, makes the same predictions:

$$\begin{aligned} Xw' &= X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\ &= Xw + \alpha(X_{1:} - X_{2:}) = Xw \end{aligned}$$

- now suppose a new person x arrives with a mutation in gene 1 ($x_1 = 1$) but not in gene 2 ($x_2 = 0$).

Q: do w and w' make the same prediction?

A. yes
B. no

Q: what criteria might you pick to choose a good w?
A: pick a w that's small; it will make less crazy predictions

Outline

Quadratic regularization

add a small penalty for large coefficients

$$\text{minimize} \quad \|y - Xw\|^2 + \lambda \|w\|^2$$

where $\lambda > 0$ is the **regularization parameter**

(also called "regularized least squares", "ridge regression", "Tikhonov regularization", or "weight decay")

why regularize?

- prevent overfitting
- stabilize estimate
- solution is always unique

Solving regularized regression

$$\text{minimize} \quad \|y - Xw\|^2 + \lambda\|w\|^2$$

- solve by setting the derivative to 0: optimal w^{ridge} satisfies

$$\begin{aligned} 0 &= \nabla^{\text{ridge}}\left(\|y - Xw^{\text{ridge}}\|^2 + \lambda\|w^{\text{ridge}}\|^2\right) \\ &= -2X^T y + 2X^T X w^{\text{ridge}} + 2\lambda w^{\text{ridge}} \\ (X^T X + \lambda I)w^{\text{ridge}} &= X^T y \end{aligned}$$

Poll: is $X^T X + \lambda I$ invertible for $\lambda > 0$?

A. always

B. if λ is larger than the smallest eigenvalue of $X^T X$

C. if X is full rank

D. never

Review: why is $X^T X + \lambda I$ invertible?

- let
$$X = U\Sigma V^T$$
be the full SVD
- then
$$\begin{aligned} X^T X + \lambda I &= V\Sigma U^T U\Sigma V^T + \lambda I \\ &= V\Sigma^2 V^T + \lambda V V^T = V(\Sigma^2 + \lambda I)V^T. \end{aligned}$$

Review: why is $X^T X + \lambda I$ invertible?

- let
$$X = U\Sigma V^T$$
be the full SVD
- then
$$\begin{aligned} X^T X + \lambda I &= V\Sigma U^T U\Sigma V^T + \lambda I \\ &= V\Sigma^2 V^T + \lambda V V^T = V(\Sigma^2 + \lambda I)V^T. \end{aligned}$$

- use the fact that for the full SVD, $V^{-1} = V^T$
- and $\Sigma^2 + \lambda I$ is diagonal with strictly positive entries, so invertible

Review: why is $X^T X + \lambda I$ invertible?

- let
$$X = U \Sigma V^T$$
be the full SVD
- then
$$\begin{aligned} X^T X + \lambda I &= V \Sigma U^T U \Sigma V^T + \lambda I \\ &= V \Sigma^2 V^T + \lambda V V^T = V(\Sigma^2 + \lambda I) V^T. \end{aligned}$$

- use the fact that for the full SVD, $V^{-1} = V^T$
- and $\Sigma^2 + \lambda I$ is diagonal with strictly positive entries, so invertible
- let's compute the inverse:
$$(X^T X + \lambda I)^{-1} = (V^T)^{-1} (\Sigma^2 + \lambda I)^{-1} V^{-1} = V(\Sigma^2 + \lambda I)^{-1} V^T.$$

Solving regularized regression

$$\text{minimize} \quad \|y - Xw\|^2 + \lambda\|w\|^2$$

- solve by setting the derivative to 0: optimal w^{ridge} satisfies

$$\begin{aligned} 0 &= \nabla^{\text{ridge}}\left(\|y - Xw^{\text{ridge}}\|^2 + \lambda\|w^{\text{ridge}}\|^2\right) \\ &= -2X^T y + 2X^T X w^{\text{ridge}} + 2\lambda w^{\text{ridge}} \\ (X^T X + \lambda I)w^{\text{ridge}} &= X^T y \end{aligned}$$

- $X^T X + \lambda I$ is **always** invertible, so

$$w^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

Quadratic regularization and the SVD

suppose $X = U\Sigma V^T$ is the (full) SVD of X.

regularized solution is

$$\begin{aligned} w^{\text{ridge}} &= (X^TX + \lambda I)^{-1}X^Ty \\ &= (V\Sigma U^TU\Sigma V^T + \lambda I)^{-1}V\Sigma U^Ty \\ &= (V\Sigma^2V^T + V(\lambda I)V^T)^{-1}V\Sigma U^Ty \\ &= V(\Sigma^2 + \lambda I)^{-1}V^TV\Sigma U^Ty \\ &= V(\Sigma^2 + \lambda I)^{-1}\Sigma U^Ty \\ &= \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y \end{aligned}$$

Quadratic regularization and the SVD

suppose $X = U\Sigma V^T$ is the (full) SVD of X.

regularized solution is

$$\begin{aligned}
w^{\text{ridge}} &= (X^TX + \lambda I)^{-1}X^Ty \\
&= (V\Sigma U^TU\Sigma V^T + \lambda I)^{-1}V\Sigma U^Ty \\
&= (V\Sigma^2V^T + V(\lambda I)V^T)^{-1}V\Sigma U^Ty \\
&= V(\Sigma^2 + \lambda I)^{-1}V^TV\Sigma U^Ty \\
&= V(\Sigma^2 + \lambda I)^{-1}\Sigma U^Ty \\
&= \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y
\end{aligned}$$

ridge regression shrinks $\sigma_i^{-1} = \frac{\sigma_i}{\sigma_i^2}$ to $\frac{\sigma_i}{\sigma_i^2+\lambda}$