ORIE 4741: Learning with Big Messy Data

Underdetermined Least Squares and Quadratic Regularization

Professor Udell
Operations Research and Information Engineering
Cornell

October 6, 2020
Announcements

- add topic: SVD for dim reduction (used in Airbnb hw)
Linear algebra review

Definition

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

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

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

The following conditions are equivalent:

- $\text{nullspace}(X) = \{ 0 \}$
- If $Xw = 0$, then $w = 0$
- The columns of $X$ are linearly independent
- $\forall z \in \mathbb{R}^n$, 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 $\mathbb{R}^d$
What if the Gram matrix is not invertible?

- Least squares objective:
  \[
  \text{minimize} \quad \| y - Xw \|^2
  \]

- Normal equations:
  \[
  X^T X w = X^T y
  \]

- Solution if \( X^T X \) is invertible:
  \[
  w = (X^T X)^{-1} X^T y
  \]

Q: if \( X^T X \) is not invertible, do the normal equations still define the solution?
What if the Gram matrix is not invertible?

- Least squares objective:
  \[
  \text{minimize } \|y - Xw\|^2
  \]

- Normal equations:
  \[
  X^TXw = X^T y
  \]

- Solution if \( X^TX \) is invertible:
  \[
  w = (X^TX)^{-1}X^T y
  \]

**Q:** if \( X^TX \) is not invertible, do the normal equations still define the solution?

**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 $X \in \mathbb{R}^{n \times d}$ in terms of easier matrices:

- $X = U \Sigma V^T$
- $U \in \mathbb{R}^{n \times d}$ is orthogonal: $U^T U = I_d$
- $V \in \mathbb{R}^{d \times d}$ is orthogonal: $V^T V = VV^T = I_d$
- $\Sigma \in \mathbb{R}^{d \times d}$ 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 \( X \in \mathbb{R}^{n \times d} \) in terms of easier matrices:

\[
X = U \Sigma V^T
\]

\( U \in \mathbb{R}^{n \times d} \) is orthogonal: \( U^T U = I_d \)

\( V \in \mathbb{R}^{d \times d} \) is orthogonal: \( V^T V = V V^T = I_d \)

\( \Sigma \in \mathbb{R}^{d \times d} \) is diagonal and nonnegative:

\[
\Sigma_{ii} \geq 0 \text{ for } i = 1, \ldots, d
\]

\[
\Sigma_{ij} = 0 \text{ for } i \neq j
\]

in julia (or matlab), use the SVD function

\[
U, S, V = \text{svd}(X)
\]

can compute \( SVD \) factorization of \( X \) in \( O(nd^2) \) flops
Thin SVD

previous version sometimes called full SVD.
to make thin SVD, delete zeros from $\Sigma$

- $r = \text{Rank}(X)$
- $X = U\Sigma V^T$
- $U \in \mathbb{R}^{n \times r}$ has orthogonal columns: $U^T U = I_r$
- $V \in \mathbb{R}^{d \times r}$ has orthogonal columns: $V^T V = I_r$
- $\Sigma \in \mathbb{R}^{r \times r}$ 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

\[
X^T Xw = X^T y \\
V\Sigma^2 V^T w = V\Sigma U^T y
\]
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

\[
X^T Xw = X^T y \\
V\Sigma^2 V^T w = V\Sigma U^T y \\
\Sigma^{-2} V^T V\Sigma^2 V^T w = \Sigma^{-2} V^T V\Sigma U^T y
\]
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

\[
\begin{align*}
X^T X w &= X^T y \\
V \Sigma^2 V^T w &= V \Sigma U^T y \\
\Sigma^{-2} V^T V \Sigma^2 V^T w &= \Sigma^{-2} V^T V \Sigma U^T y \\
V^T w &= \Sigma^{-1} U^T y
\end{align*}
\]

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

\[
X^T X w = X^T y \\
V \Sigma^2 V^T w = V \Sigma U^T y \\
\Sigma^{-2} V^T V \Sigma^2 V^T w = \Sigma^{-2} V^T V \Sigma U^T y \\
V^T w = \Sigma^{-1} U^T y
\]

can’t solve (\( V^T \) not invertible, solution not unique...) try

\[
w = V \Sigma^{-1} U^T y = \sum_{i=1}^{d} v_i \sigma_i^{-1} u_i^T y:
\]
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
\[
\begin{align*}
X^T X w &= X^T y \\
V\Sigma^2 V^T w &= V\Sigma U^T y \\
\Sigma^{-2} V^T V\Sigma^2 V^T w &= \Sigma^{-2} V^T V\Sigma U^T y \\
V^T w &= \Sigma^{-1} U^T y
\end{align*}
\]

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

so we’ve found a solution (without assuming invertibility)!
Demo: SVD

https://github.com/ORIE4741/demos/SVD.ipynb
Review: methods for least squares

<table>
<thead>
<tr>
<th></th>
<th>GD</th>
<th>SGM</th>
<th>Gram GD</th>
<th>Parallel GD</th>
<th>QR or SVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial per iter</td>
<td>0</td>
<td>0</td>
<td>$nd^2$</td>
<td>$nd^2/P$</td>
<td>$nd^2$</td>
</tr>
<tr>
<td></td>
<td>$nd$</td>
<td>$</td>
<td>S</td>
<td>d$</td>
<td>$d^2$</td>
</tr>
</tbody>
</table>

(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)
Outline

The SVD

Non-uniqueness

Quadratic regularization
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]

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

A: no, if \( X^T X \) is not invertible, the solution is not unique! if \( \text{Rank}(X^T X) < d \), 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 \( \alpha \in \mathbb{R} \).

Q: is non-uniqueness a problem for a predictive model?

A: yes.
What if the Gram matrix is not invertible?

\[ X^T Xw = X^T y \]

**Q:** is the solution to the normal equations always unique?  
**A:** no, if \( X^T X \) is not invertible, the solution is not unique!  
If \( \text{Rank}(X^T X) < d \), then for some \( v \neq 0 \), \( X^T Xv = 0 \).  
So if \( X^T Xw = X^T y \), then \( X^T X(w + \alpha v) = X^T y \) for any \( \alpha \in \mathbb{R} \).
What if the Gram matrix is not invertible?

\[ X^T Xw = X^T y \]

Q: is the solution to the normal equations always unique?
A: no, if \( X^T X \) is not invertible, the solution is not unique!
if \( \text{Rank}(X^T X) < d \), then for some \( v \neq 0 \), \( X^T Xv = 0 \).
so if \( X^T Xw = X^T y \), then \( X^T X(w + \alpha v) = X^T y \) for any \( \alpha \in \mathbb{R} \).

Q: is non-uniqueness a problem for a predictive model?
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]

**Q:** is the solution to the normal equations always unique?  
**A:** no, if \( X^T X \) is not invertible, the solution is not unique!  
if \( \text{Rank}(X^T X) < d \), 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 \( \alpha \in \mathbb{R} \).

**Q:** is non-uniqueness a problem for a predictive model?  
**A:** yes.
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 \( \alpha \in \mathbb{R} \), makes the same predictions:
  \[
  Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2)
  = Xw + \alpha (X_1: - X_2:) = Xw
  \]

- 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 \( \alpha \in \mathbb{R} \), makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2)
= Xw + \alpha(X_{1:} - X_{2:}) = Xw
\]

- 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?
Example: non-uniqueness (II)

\[ X_1: = X_2: \]

- Suppose our least squares solution is \( w \).
- \( w' = w + \alpha e_1 - \alpha e_2 \), for \( \alpha \in \mathbb{R} \), makes the same predictions:
  \[ Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \]
  \[ = Xw + \alpha (X_1: - X_2:) = Xw \]

- 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:** 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 \( \alpha \in \mathbb{R} \), makes the same predictions:

\[
Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\
= Xw + \alpha(X_1: - X_2:) = Xw
\]

- 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: no!
Q: what criterion 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 \( \alpha \in \mathbb{R} \), makes the same predictions:
  \[
  Xw' = X(w + \alpha e_1 - \alpha e_2) = Xw + \alpha X(e_1 - e_2) \\
  = Xw + \alpha (X_1: - X_2:) = Xw
  \]

- 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: no!

Q: what criterion might you pick to choose a good \( w \)?
A: pick a \( w \) that’s small; it will make less crazy predictions
Outline

The SVD

Non-uniqueness

Quadratic regularization
add a small penalty for large coefficients

\[
\text{minimize } \| y - Xw \|^2 + \lambda \| w \|^2
\]

where \( \lambda > 0 \) is the \textbf{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

\[
\begin{align*}
\text{minimize} & \quad \| y - Xw \|^2 + \lambda \| w \|^2 \\
\text{solve by setting the derivative to 0: optimal } w^{\text{ridge}} & \text{ satisfies} \\
0 & = \nabla^{\text{ridge}} \left( \| y - Xw^{\text{ridge}} \|^2 + \lambda \| w^{\text{ridge}} \|^2 \right) \\
& = -2X^Ty + 2X^TXw^{\text{ridge}} + 2\lambda w^{\text{ridge}} \\
(X^TX + \lambda I)w^{\text{ridge}} & = X^Ty \\
\end{align*}
\]

\[\begin{align*}
X^TX + \lambda I & \text{ is always invertible, so} \\
w^{\text{ridge}} & = (X^TX + \lambda I)^{-1}X^Ty
\end{align*}\]
Review: why is $X^TX + \lambda I$ invertible?

- let

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

be the full SVD

- then

$$X^TX + \lambda I = V\Sigma U^T U\Sigma V^T + \lambda I = V\Sigma^2 V^T + \lambda VV^T = V(\Sigma^2 + \lambda I)V$$
Review: why is $X^T X + \lambda I$ invertible?

- let
  \[ X = U \Sigma V^T \]
  be the full SVD

- then
  \[ X^T X + \lambda I = V \Sigma U^T U \Sigma V^T + \lambda I = V \Sigma^2 V^T + \lambda V V^T = V(\Sigma^2 + \lambda I) V \]

- 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 $X^TX + \lambda I$ invertible?

▸ let

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

be the full SVD

▸ then

$$X^TX + \lambda I = V\Sigma U^TU\Sigma V^T + \lambda I = V\Sigma^2 V^T + \lambda VV^T = V(\Sigma^2 + \lambda I)V^T.$$  

▸ 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:

$$(X^TX + \lambda I)^{-1} = (V^T)^{-1}(\Sigma^2 + \lambda I)^{-1}V^{-1} = V(\Sigma^2 + \lambda I)^{-1}V^T.$$
Quadratic regularization and the SVD

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

regularized solution is

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

$$ = (V \Sigma U^T U \Sigma V^T + \lambda I)^{-1} V \Sigma U^T y $$

$$ = (V \Sigma^2 V^T + V(\lambda I)V^T)^{-1} V \Sigma U^T y $$

$$ = V(\Sigma^2 + \lambda I)^{-1} V^T V \Sigma U^T y $$

$$ = V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y $$

$$ = \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y $$

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