**Linear algebra review**

**Definition**

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

$$X = \{ w \in \mathbb{R}^d : Xw = 0 \}$$

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

The following conditions are equivalent:

- $(X) = \{0\}$
- 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 $\mathbb{R}^d$
Outline

Motivation

The SVD

Regularizing underdetermined problems

Why regularization helps
What if the Gram matrix is not invertible?

\[ X^T Xw = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

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

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

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

1. **Q:** if $X^T X$ is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it's not unique!

   If $\text{Rank}(X) < d$, then for some $v \neq 0$, $Xv = 0$.

   So if $w$ is a solution, so is $w + \alpha v$ for any $\alpha \in \mathbb{R}$.

3. **Q:** is non-uniqueness a problem for a predictive model?
   **A:** yes.
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?

   If \( \text{Rank}(X^T X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).

   So if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).

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

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

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it’s not unique!
   if \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).
   so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).
What if the Gram matrix is not invertible?

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it’s not unique!
   if \( \text{Rank}(X) < d \), then for some \( v \neq 0 \), \( Xv = 0 \).
   so if \( w \) is a solution, so is \( w + \alpha v \) for any \( \alpha \in \mathbb{R} \).

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

\[ X^T X w = X^T y \]
\[ w = (X^T X)^{-1} X^T y \]

1. **Q:** if \( X^T X \) is not invertible, do the normal equations still define the solution?
   **A:** yes! we derived them with no assumptions.

2. **Q:** if they have a solution, is the solution unique?
   **A:** no, it’s not unique!
   if \( \text{Rank}(X) < d \), then for some \( \nu \neq 0 \), \( X \nu = 0 \).
   so if \( w \) is a solution, so is \( w + \alpha \nu \) for any \( \alpha \in \mathbb{R} \).

3. **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

Motivation

The SVD

Regularizing underdetermined problems

Why regularization helps
The Singular Value Decomposition (SVD)

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

- $X = U \Sigma V^T$
- $U \in \mathbb{R}^{n \times r}$ 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 $i \neq j$

can compute SVD factorization of $X$ in $O(nd^2)$ flops
The Singular Value Decomposition (SVD)

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

- $X = U \Sigma V^T$
- $U \in \mathbb{R}^{n \times r}$ 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 $i \neq j$

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

in julia (or matlab), use the SVD function

$$ U, S, V = \text{svd}(X) $$
Full SVD

previous version sometimes called thin SVD.
to make full SVD, augment Σ with zeros.
suppose $d \leq n$. full SVD is

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

$V$ is square and orthogonal, so $V^T V = V V^T = I_d$
SVD for least squares

if \( X = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \), then

\[
X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T
\]

normal equations are

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

try \( w = V\Sigma^{-1} U^T y = \sum_{i=1}^{d} v_i \frac{1}{\sigma_i} u_i^T y \):

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

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

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

- 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 for underdetermined problems, $O(nd^2)$ flops)
Outline

Motivation

The SVD

Regularizing underdetermined problems

Why regularization helps
**Quadratic regularization**

add a small penalty for large coefficients

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

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

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

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

\[\begin{align*}
X^TX + \lambda I \text{ is always invertible, so} \\
w^{\text{ridge}} = (X^TX + \lambda I)^{-1}X^Ty \\
\end{align*}\]
suppose $X = UΣV^T$ is the (full) SVD of $X$. regularized solution is

$$w^{\text{ridge}} = (X^TX + λI)^{-1}VΣU^Ty$$

$$= (VΣ^2V^T + V(λI)V^T)^{-1}VΣU^Ty$$

$$= V(Σ^2 + λI)^{-1}V^TVΣU^Ty$$

$$= V(Σ^2 + λI)^{-1}ΣU^Ty$$

$$= \sum_{i=1}^{d} \nu_i \frac{σ_i}{σ_i^2 + λ} u_i^Ty$$

ridge regression shrinks $σ_i^{-1} = \frac{σ_i}{σ_i^2}$ to $\frac{σ_i}{σ_i^2 + λ}$
Outline

Motivation

The SVD

Regularizing underdetermined problems

Why regularization helps
Bias variance tradeoff for regression

- Suppose $y = Xw^\dagger + \epsilon$
- $X = U\Sigma V^T$ is the SVD of $X$

$$w^{\text{ridge}} = \sum_{i=1}^{d} v_i \frac{\sigma_i}{\sigma_i^2 + \lambda} u_i^T y, \quad w^{\text{lsq}} = \sum_{i=1}^{d} v_i \frac{1}{\sigma_i} u_i^T y$$
Bias variance tradeoff: least squares regression

- Suppose $y = Xw^\dagger + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$
- Different samples of datasets $D$ have same $X$, different $\varepsilon$
- $X = U\Sigma V^T$ is the SVD of $X$

\[
\begin{align*}
    f(x) &= x^T w^\dagger \\
    g_D(x) &= x^T (X^TX)^{-1} X^Ty = x^T (X^TX)^{-1} X^T (Xw^\dagger + \varepsilon) \\
    &= x^T w^\dagger + x^T (X^TX)^{-1} X^T \varepsilon \\
    \bar{g}(x) &= \mathbb{E}_D [g_D(x)] = x^T w^\dagger
\end{align*}
\]
Bias variance tradeoff: least squares regression

so

\[ \text{bias}(x) = f(x) - \bar{g}(x) = 0 \]
\[ \text{var}(x) = \mathbb{E}_D \left[ (g_D(x) - \bar{g}(x))^2 \right] \]
\[ = \mathbb{E}_D \left[ x^T (X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1} x \right] \]
\[ = x^T (X^T X)^{-1} X^T \mathbb{E}_D \left[ \varepsilon \varepsilon^T \right] X (X^T X)^{-1} x \]
\[ = x^T (X^T X)^{-1} X^T I X (X^T X)^{-1} x \]
\[ = x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x \]
\[ = x^T (X^T X)^{-1} x \]
\[ = x^T \left( \sum_{i=1}^{d} \frac{1}{\sigma_i^2} v_i v_i^T \right) x \]
Bias variance tradeoff: ridge regression

- Suppose $y = Xw^\dagger + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, n$
- Different samples of datasets $D$ have the same $X$, different $\varepsilon$
- $X = U\Sigma V^T$ is the SVD of $X$

\[
\begin{align*}
  f(x) &= x^T w^\dagger \\
  g_D(x) &= x^T w^{\text{ridge}} = x^T (X^T X + \lambda I)^{-1} X^T y \\
  &= x^T (X^T X + \lambda I)^{-1} X^T (Xw^\dagger + \varepsilon) \\
  \bar{g}(x) &= \mathbb{E}_D [g_D(x)] = x^T (X^T X + \lambda I)^{-1} X^T Xw^\dagger
\end{align*}
\]
Bias variance tradeoff: ridge regression

so

\[ \text{bias}(x) = f(x) - \bar{g}(x) = x^T((X^TX + \lambda I)^{-1}X^TX - I)w_{\parallel} \]

\[ \text{var}(x) = \mathbb{E}_D[(g_D(x) - \bar{g}(x))^2] \]

\[ = \mathbb{E}_D\left[x^T(X^TX + \lambda I)^{-1}X^T\varepsilon\varepsilon^TX(X^TX + \lambda I)^{-1}x\right] \]

\[ = x^T(X^TX + \lambda I)^{-1}X^T\mathbb{E}_D[\varepsilon\varepsilon^T]X(X^TX + \lambda I)^{-1}x \]

\[ = x^T(X^TX + \lambda I)^{-1}X^TIX(X^TX + \lambda I)^{-1}x \]

\[ = x^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}x \]

\[ = x^T \left( \sum_{i=1}^{d} v_i \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} v_i^T \right) x \]