Regularization

**Linear regression when \( d > n \)**

Data in matrix/vector form

\[
A := \begin{bmatrix}
\leftarrow x_1^\top \\
\leftarrow x_2^\top \\
\vdots \\
\leftarrow x_n^\top 
\end{bmatrix}_{n \times d \text{ matrix}}, \quad b := \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n 
\end{bmatrix}_{n \times 1 \text{ vector}}.
\]

Ordinary least squares

Ordinary least squares \( \hat{w}_{\text{ols}} \): typically “defined” by \( \hat{w}_{\text{ols}} := (A^\top A)^{-1} A^\top b \).

Ill-defined when \( d > n \). In this case, the \( d \times d \) matrix

\[
A^\top A = \sum_{i=1}^{n} x_i x_i^\top
\]

is not invertible (as its rank is at most \( n < d \)).

**Regularization**

Typical solution: regularization

Some examples:

- encourage \( \|w\|_2^2 \) to be small (“ridge regression”)
- encourage \( \|w\|_1 \) to be small (“Lasso”)
- encourage \( w \) to be sparse (“sparse regression”)

Example: regularized least squares criterion

Find \( w \in \mathbb{R}^d \) to minimize

\[
\|A w - b\|_2^2 + \lambda R(w)
\]

where \( \lambda > 0 \) and \( R: \mathbb{R}^d \rightarrow \mathbb{R}_+ \) is a penalty function / regularizer.

**Understanding regularization**

How do I decide which type of regularization to use?

- Could just try them all . . .
- Better answer: try to understand their statistical behavior in a broad class of scenarios.
RIDGE REGRESSION

Find $w \in \mathbb{R}^d$ to minimize

$$\|Aw - b\|_2^2 + \lambda \|w\|_2^2$$

where $\lambda > 0$.

This always has a unique solution.

RIDGE REGRESSION VIA CALCULUS

Ridge regression objective is convex function of $w$.
Suffices to find $w$ where gradient is zero.

$$\nabla_w \left\{ \|Aw - b\|_2^2 + \lambda \|w\|_2^2 \right\} = 2A^\top (Aw - b) + 2\lambda w.$$  

This is zero when

$$(A^\top A + \lambda I)w = A^\top b,$$

a system of linear equations in $w$.
Matrix $A^\top A + \lambda I$ is invertible since $\lambda > 0$, so its unique solution is

$$\hat{w}_{\text{ridge}} := (A^\top A + \lambda I)^{-1} A^\top b.$$  

RIDGE REGRESSION: GEOMETRY

If $\hat{w}_{\text{ols}}$ exists, then ridge regression objective (as function of $w$) is

$$\|A(w - \hat{w}_{\text{ols}})\|_2^2 + \lambda \|w\|_2^2 + \text{(stuff not depending on } w),$$

Level sets for $\|A(w - \hat{w}_{\text{ols}})\|_2^2$

Level sets for $\lambda \|w\|_2^2$
Aside: Eigendecompositions

Every symmetric matrix $M \in \mathbb{R}^{d \times d}$ guaranteed to have eigendecomposition with real eigenvalues:

$$M = V \Lambda V^\top$$

**real eigenvalues:** $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ ($\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$);

**corresponding orthonormal eigenvectors:** $v_1, v_2, \ldots, v_d$ ($V = [v_1 \mid v_2 | \cdots | v_d]$).

Eigenvectors $v_1, v_2, \ldots, v_d$ constitute an orthonormal basis for $\mathbb{R}^d$.

So every $w \in \mathbb{R}^d$ can be written as a linear combination of these vectors:

$$w = \sum_{j=1}^{d} \langle v_j, w \rangle v_j.$$

Ridge regression: eigendecomposition

Write eigendecomposition of $A^\top A$ as

$$A^\top A = \sum_{j=1}^{d} \lambda_j v_j v_j^\top$$

where $v_1, v_2, \ldots, v_d \in \mathbb{R}^d$ are orthonormal eigenvectors with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$.

- For $\lambda > 0$, the inverse of $A^\top A + \lambda I$ exists, and has the form

$$\left(A^\top A + \lambda I\right)^{-1} = \sum_{j=1}^{d} \frac{1}{\lambda_j + \lambda} v_j v_j^\top.$$

Ridge regression vs. ordinary least squares

If $\hat{w}_{\text{ols}}$ exists, then

$$\hat{w}_{\text{ridge}} = \left(A^\top A + \lambda I\right)^{-1} \left(A^\top A\right) \hat{w}_{\text{ols}}$$

$$= \left(\sum_{j=1}^{d} \frac{1}{\lambda_j + \lambda} v_j v_j^\top\right) \left(\sum_{j=1}^{d} \lambda_j v_j v_j^\top\right) \hat{w}_{\text{ols}}$$

$$= \left(\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda} v_j v_j^\top\right) \hat{w}_{\text{ols}} \quad \text{(by orthogonality)}$$

$$= \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda} \langle v_j, \hat{w}_{\text{ols}} \rangle v_j.$$

**Interpretation:** Shrink $\hat{w}_{\text{ols}}$ towards zero by $\frac{\lambda_j}{\lambda_j + \lambda}$ factor in direction $v_j$.

**Effective degrees-of-freedom:** $\text{df}(\lambda) := \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$.

Coefficient profile

Horizontal axis: varying $\lambda$ (large $\lambda$ to left, small $\lambda$ to right).  
Vertical axis: coefficient value in $\hat{w}_{\text{ridge}}$ for eight different variables.
fixed-design setting

Easier/cleaner to study OLS and ridge regression in the **fixed design** setting:

- Assume $x_1, x_2, \ldots, x_n$ are not random; only $y_1, y_2, \ldots, y_n$ are random. ($A$ is not random; only $b$ is random.)
- Best predictor of $y_i$ is $E[y_i]$ (in terms of expected square loss).
  Note: $x_i \mapsto E[y_i]$ might not be realized by a linear function.
- Let $w_\star \in \mathbb{R}^d$ be a weight vector that minimizes
  \[ w \mapsto E[\|Aw - b\|^2]. \]
  (best linear predictor).
  If $A^\top A$ is invertible, then $w_\star = (A^\top A)^{-1} A^\top E[b]$.
- **Basic question:** how well can we estimate $w_\star$?

OLS: fixed-design analysis

When $\hat{w}_{ols}$ exists (i.e., when $A^\top A$ is invertible), then it is an unbiased estimator of $w_\star$:

\[ E[\hat{w}_{ols}] = w_\star. \]

Let $\varepsilon := b - Aw_\star$. Then

\[ \text{cov}(\hat{w}_{ols}) = (A^\top A)^{-1} A^\top E[\varepsilon \varepsilon^\top] A(A^\top A)^{-1}. \]

For example, if $\varepsilon \sim N(0, \sigma^2 I)$, then

\[ \text{cov}(\hat{w}_{ols}) = \sigma^2 (A^\top A)^{-1} = \sigma^2 \sum_{j=1}^d \frac{1}{\lambda_j} v_j v_j^\top, \]

so the variance of $\langle v_j, \hat{w}_{ols} \rangle$ is

\[ \text{var}(\langle v_j, \hat{w}_{ols} \rangle) = v_j^\top \text{cov}(\hat{w}_{ols}) v_j = \frac{\sigma^2}{\lambda_j}. \]

Note: if $\lambda_d$ is very close to zero (so $A^\top A$ is close to being singular), then the variance in direction $v_d$ is very high.
RIDGE REGRESSION: FIXED-DESIGN ANALYSIS

Ridge regression is not an unbiased estimator $w_*$:

$$\mathbb{E}[\hat{w}_{\text{ridge}}] \neq w_*.$$  

But, covariance of $\hat{w}_{\text{ridge}}$ is always “smaller” than that of $\hat{w}_{\text{ols}}$:

$$\text{cov}(\hat{w}_{\text{ridge}}) = (A^T A + \lambda I)^{-1} A^T \mathbb{E}[\epsilon \epsilon^T] A (A^T A + \lambda I)^{-1}$$

$$= M \text{cov}(\hat{w}_{\text{ols}}) M \quad \text{(if } \hat{w}_{\text{ols}} \text{ exists)},$$

where

$$M := \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \lambda} v_j v_j^T.$$  

For example, if $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, then

$$\text{cov}(\hat{w}_{\text{ridge}}) = \sigma^2 \sum_{j=1}^d \frac{\lambda_j}{(\lambda_j + \lambda)^2} v_j v_j^T = \sigma^2 \sum_{j: \lambda_j > 0} \left(1 + \frac{\lambda_j}{\lambda_j + \lambda} \right) \frac{1}{\lambda_j} v_j v_j^T.$$  

Bias-variance trade-off

Very explicit bias-variance trade-off

$$w_* - \mathbb{E}[\hat{w}_{\text{ridge}}] = \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \lambda} (v_j, w_*) v_j,$$

$$\text{cov}(\hat{w}_{\text{ridge}}) = (\lambda^T A + \lambda I)^{-1} A^T \mathbb{E}[\epsilon \epsilon^T] A (A^T A + \lambda I)^{-1}.$$  

Using a similar analysis, can also reveal trade-off in excess expected square loss:

$$\mathbb{E}\left[\|A \hat{w}_{\text{ridge}} - b\|_2^2\right] - \mathbb{E}\left[\|A w_* - b\|_2^2\right].$$

OTHER INTERPRETATIONS

- Suppose we replace $A$ and $b$ with

$$\tilde{A} := \begin{bmatrix} A \\ \sqrt{\lambda} Q \end{bmatrix}, \quad \tilde{b} := \begin{bmatrix} b \\ 0 \end{bmatrix},$$

where $Q \in \mathbb{R}^{(n+d) \times d}$ is any orthogonal matrix. That is, add $d$ fictitious labeled data $(\sqrt{\lambda} q_j, 0)$ for $j = 1, 2, \ldots, d$, where $q_1, q_2, \ldots, q_d$ is an orthonormal basis.

- Then $\tilde{A}^T \tilde{A} = A^T A + \lambda I$, and $\tilde{A}^T \tilde{b} = A^T b$.

- So $\tilde{w}_{\text{ols}}$ using $\tilde{A}$ and $\tilde{b}$ is the same as $\hat{w}_{\text{ridge}}$ using $A$ and $b$.

- For a similar reason, $\hat{w}_{\text{ridge}}$ has a certain Bayesian interpretation.

RIDGE REGRESSION: SUMMARY

- Ridge regression always well-defined (for $\lambda > 0$), unlike OLS (which is degenerate ridge regression with $\lambda = 0$).

- Behavior depends on eigenvectors/eigenvalues of $A^T A$; the original “coordinate system” is not relevant here.

- Relative to $\hat{w}_{\text{ols}}$: shrinks $\hat{w}_{\text{ols}}$ along eigenvector directions by amount related to eigenvalue and $\lambda$.

- Regularization parameter $\lambda$ is tuning-knob that controls bias-variance trade-off.

- Can be thought of as applying OLS to an augmented data set with “fake data” that ensures OLS is well-defined.
**Sparse regression**

Another form of regularization: only consider sparse $w$—i.e., $w$ with only a small number ($\ll d$) of non-zero entries.

Other advantages of sparsity (especially relative to ridge):
- Sparse solutions easier to “interpret” (but caveats about interpreting weights from before still apply).
- Can be more efficient to evaluate $\langle w, x \rangle$ (both in terms of computing variable values and computing inner product).

**Sparse regression methods**

For any $T \subseteq \{1, 2, \ldots, d\}$, let $\hat{w}_T :=$ OLS only using variables in $T$.

**Subset selection**

**Brute-force strategy.** Pick the $T \subseteq \{1, 2, \ldots, d\}$ of size $|T| = k$ for which

$$\|A\hat{w}_T - b\|_2^2$$

is minimal, and return $\hat{w}_T$.

- Gives you exactly what you want (for given value $k$).
- Only feasible for very small $k$, since complexity scales with $\binom{d}{k}$.
- (NP-hard optimization problem.)

**Forward stepwise regression**

**Greedy strategy.** Starting with $T = \emptyset$, repeat until $|T| = k$:

Pick the $j \in \{1, 2, \ldots, d\} \setminus T$ for which

$$\|A\hat{w}_{T \cup \{j\}} - b\|_2^2$$

is minimal, and add this $j$ to $T$.

Return $\hat{w}(T)$.

- Gives you a $k$-sparse solution.
- Primarily only effective when columns of $A$ are close to orthogonal.
Aside: $l_p$ norms

For $p \geq 1$, 
\[
\|v\|_p = \left( \sum_{j=1}^{d} |v_j|^p \right)^{1/p}.
\]
In particular, 
\[
\|v\|_1 = \sum_{j=1}^{d} |v_j|.
\]
These are norms on $\mathbb{R}^d$, so:
- $\|u - v\|_p$ is a valid metric on points in $\mathbb{R}^d$, and
- $\|cv\|_p = |c| \cdot \|v\|_p$ for any $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$.

Lasso: least absolute shrinkage and selection operator

Let $\hat{w}_{\text{lasso}}$ be a minimizer of
\[
\arg \min_{w \in \mathbb{R}^p} \|A w - b\|_2^2 + \lambda \|w\|_1.
\]
Objective function is convex though not differentiable.

If $\hat{w}_{\text{ols}}$ exists, then Lasso objective (as function of $w$) is
\[
\|A(w - \hat{w}_{\text{ols}})\|_2^2 + \lambda \|w\|_1 + \text{(stuff not depending on $w$)}.
\]

Coefficient profile (Ridge vs. Lasso)

Horizontal axis: varying $\lambda$ (large $\lambda$ to left, small $\lambda$ to right).
Vertical axis: coefficient value in $\hat{w}_{\text{ridge}}$ and $\hat{w}_{\text{lasso}}$ for eight different variables.

Lasso: theory

Many results, mostly roughly of the following flavor.
Suppose
- $b \sim N(Aw_*, \sigma^2 I)$;
- $w_*$ has $\leq k$ non-zero entries;
- $A$ satisfies some special properties (typically not efficiently checkable);
- $\lambda \gtrsim \sigma \sqrt{2n \log(d)}$;
then
\[
\mathbb{E}[\|w_* - \hat{w}_{\text{lasso}}\|_2^2] \leq O\left(\frac{\sigma^2 k \log(d)}{n}\right).
\]
Closely related to “compressed sensing”; theory involves high-dimensional convex geometry.
Sparse regression: summary

- **Sparsity**: a form of "regularization", but also desirable for other reasons.
- **Subset selection**: generally intractable.
- **Greedy algorithms** (e.g., forward stepwise regression): sometimes works.
- **Lasso**: shrink coefficients towards zero in a way that tends to lead to sparse solutions.