Regularization

Linear regression when $d > n$

Data in matrix/vector form

$$A := \underbrace{\begin{bmatrix} \longleftarrow & x_1^\top & \longrightarrow \\ \longleftarrow & x_2^\top & \longrightarrow \\ & \vdots & \\ \longleftarrow & x_n^\top & \longrightarrow \end{bmatrix}}_{n \times d \text{ matrix}}, \qquad b := \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1 \text{ vector}}.$$

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

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

where $\lambda > 0$ and $R \colon \mathbb{R}^d \to \mathbb{R}_+$ 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

Ridge regression

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

$$\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

where $\lambda > 0$.

This *always* has a unique solution.

Ridge regression via calculus

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

$$\nabla_{\boldsymbol{w}}\left\{\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2\right\} \;=\; 2\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}) + 2\lambda\boldsymbol{w}\,.$$

This is zero when

$$(\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})\boldsymbol{w} \;=\; \boldsymbol{A}^\top\boldsymbol{b}\,,$$

a system of linear equations in $\boldsymbol{w}$.

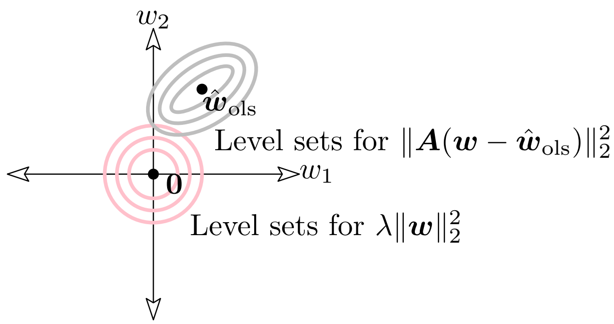
Matrix $\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I}$ is invertible since $\lambda > 0$, so its *unique* solution is

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

Ridge regression: geometry

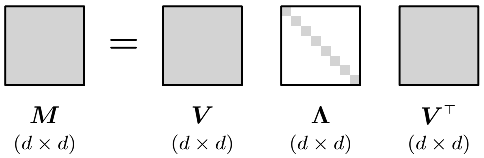
If $\hat{\boldsymbol{w}}_{\text{ols}}$ exists, then ridge regression objective (as function of $\boldsymbol{w}$) is

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



Aside: Eigendecompositions

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



$$= \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

real **eigenvalues**: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \quad (\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d))$;
corresponding orthonormal **eigenvectors**: $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d \quad (\boldsymbol{V} = [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_d])$.

Eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$ constitute an **orthonormal basis** for $\mathbb{R}^d$.

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

$$\boldsymbol{w} = \sum_{j=1}^{d} \langle \boldsymbol{v}_j, \boldsymbol{w} \rangle \boldsymbol{v}_j .$$

Ridge regression: eigendecomposition

Write eigendecomposition of $\boldsymbol{A}^\top \boldsymbol{A}$ as

$$\boldsymbol{A}^\top \boldsymbol{A} = \sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d \in \mathbb{R}^d$ are orthonormal eigenvectors with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$.

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

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

Ridge regression vs. ordinary least squares

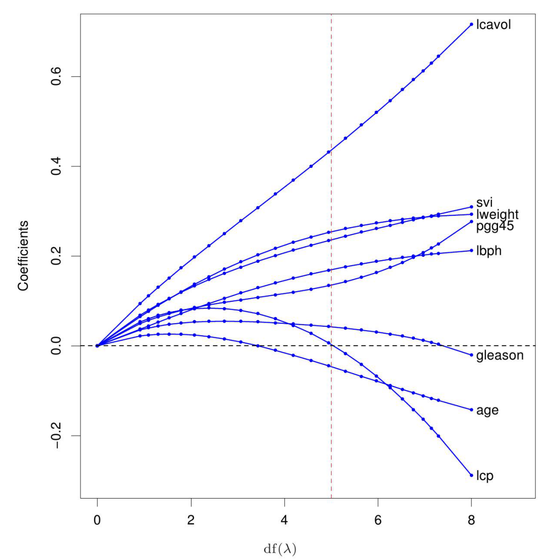
If $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then

$$\begin{aligned}
\hat{\boldsymbol{w}}_{\mathrm{ridge}} &= \left(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I} \right)^{-1} \left(\boldsymbol{A}^\top \boldsymbol{A} \right) \hat{\boldsymbol{w}}_{\mathrm{ols}} \\
&= \left(\sum_{j=1}^{d} \frac{1}{\lambda_j + \lambda} \boldsymbol{v}_j \boldsymbol{v}_j^\top \right) \left(\sum_{j=1}^{d} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top \right) \hat{\boldsymbol{w}}_{\mathrm{ols}} \\
&= \left(\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda} \boldsymbol{v}_j \boldsymbol{v}_j^\top \right) \hat{\boldsymbol{w}}_{\mathrm{ols}} \quad \text{(by orthogonality)} \\
&= \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda} \langle \boldsymbol{v}_j, \hat{\boldsymbol{w}}_{\mathrm{ols}} \rangle \boldsymbol{v}_j .
\end{aligned}$$

Interpretation: Shrink $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ towards zero by $\frac{\lambda_j}{\lambda_j + \lambda}$ factor in direction $\boldsymbol{v}_j$.

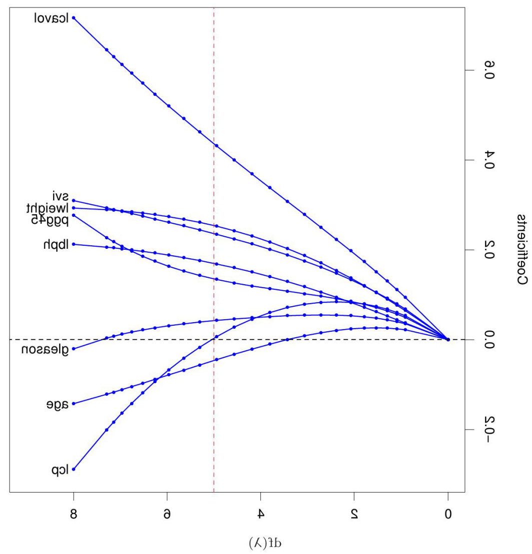
Effective degrees-of-freedom: $\operatorname{df}(\lambda) := \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda}$.

Coefficient profile



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

Coefficient profile (flipped)



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

How to compare OLS and ridge?

- **Case 1**: $\boldsymbol{A}^\top\boldsymbol{A}$ not invertible.
  No comparison, because OLS even doesn't exist.
- **Case 2**: $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible (but perhaps close to being singular).
  How do ridge regression and OLS compare?

  Is there a general setting in which to analyze OLS and ridge regression?
  - **Statistical learning setting**: data comes from unknown distribution P over $\mathbb{R}^d \times \mathbb{R}$, goal is to minimize expected square loss.
    (In context of regression, typically called **random design**.)
  - ...

Fixed-design setting

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

- Assume $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ are not random; only $y_1, y_2, \dots, y_n$ are random.
  ($\boldsymbol{A}$ is not random; only $\boldsymbol{b}$ is random.)

  Also assume all y_i have finite variance.
- Best predictor of y_i is $\mathbb{E}[y_i]$ (in terms of expected square loss).
  Note: $\boldsymbol{x}_i \mapsto \mathbb{E}[y_i]$ might not be realized by a linear function.
- Let $\boldsymbol{w}_\star \in \mathbb{R}^d$ be a weight vector that minimizes

$$\boldsymbol{w} \;\mapsto\; \mathbb{E}\left[\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2\right]$$

  (best *linear* predictor).
  If $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible, then $\boldsymbol{w}_\star = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\,\mathbb{E}[\boldsymbol{b}]$.
- **Basic question**: how well can we estimate $\boldsymbol{w}_\star$?

OLS: fixed-design analysis

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

$$\mathbb{E}[\hat{\boldsymbol{w}}_{\text{ols}}] \;=\; \boldsymbol{w}_\star\,.$$

Let $\boldsymbol{\varepsilon} := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}_\star$. Then

$$\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ols}}) \;=\; (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\,\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\,.$$

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

$$\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ols}}) \;=\; \sigma^2(\boldsymbol{A}^\top\boldsymbol{A})^{-1} \;=\; \sigma^2\sum_{j=1}^d \frac{1}{\lambda_j}\boldsymbol{v}_j\boldsymbol{v}_j^\top\,,$$

so the variance of $\langle \boldsymbol{v}_j, \hat{\boldsymbol{w}}_{\text{ols}}\rangle$ is

$$\operatorname{var}(\langle \boldsymbol{v}_j, \hat{\boldsymbol{w}}_{\text{ols}}\rangle) \;=\; \boldsymbol{v}_j^\top\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ols}})\boldsymbol{v}_j \;=\; \frac{\sigma^2}{\lambda_j}\,.$$

Note: if λ_d is very close to zero (so $\boldsymbol{A}^\top\boldsymbol{A}$ is close to being singular), then the variance in direction $\boldsymbol{v}_d$ is very high.

Ridge regression: fixed-design analysis

Ridge regression is not an *unbiased* estimator $\boldsymbol{w}_\star$:

$$\mathbb{E}[\hat{\boldsymbol{w}}_{\text{ridge}}] \neq \boldsymbol{w}_\star .$$

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

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ridge}}) &= (\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^\top\,\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})^{-1} \\ &= \boldsymbol{M}\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ols}})\boldsymbol{M} \quad (\text{if } \hat{\boldsymbol{w}}_{\text{ols}} \text{ exists}),\end{aligned}$$

where

$$\boldsymbol{M} := \sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j+\lambda}\boldsymbol{v}_j\boldsymbol{v}_j^\top .$$

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

$$\operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ridge}}) = \sigma^2\sum_{j=1}^{d}\frac{\lambda_j}{(\lambda_j+\lambda)^2}\boldsymbol{v}_j\boldsymbol{v}_j^\top = \sigma^2\sum_{j:\lambda_j>0}\underbrace{\left(\frac{1}{(1+\lambda/\lambda_j)^2}\right)}_{\leq 1}\cdot\frac{1}{\lambda_j}\boldsymbol{v}_j\boldsymbol{v}_j^\top .$$

Bias-variance trade-off

Very explicit bias-variance trade-off

$$\begin{aligned}\boldsymbol{w}_\star - \mathbb{E}[\hat{\boldsymbol{w}}_{\text{ridge}}] &= \sum_{j=1}^{d}\frac{\lambda}{\lambda_j+\lambda}\langle\boldsymbol{v}_j,\boldsymbol{w}_\star\rangle\boldsymbol{v}_j , \\ \operatorname{cov}(\hat{\boldsymbol{w}}_{\text{ridge}}) &= (\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^\top\,\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})^{-1} .\end{aligned}$$

λ	‖bias‖	variance
$\uparrow$	$\uparrow$	$\downarrow$
$\downarrow$	$\downarrow$	$\uparrow$

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

$$\mathbb{E}\Big[\|\boldsymbol{A}\hat{\boldsymbol{w}}_{\text{ridge}}-\boldsymbol{b}\|_2^2\Big] - \mathbb{E}\Big[\|\boldsymbol{A}\boldsymbol{w}_\star-\boldsymbol{b}\|_2^2\Big] .$$

Other interpretations

- Suppose we replace $\boldsymbol{A}$ and $\boldsymbol{b}$ with

$$\tilde{\boldsymbol{A}} := \underbrace{\begin{bmatrix}\boldsymbol{A}\\ \sqrt{\lambda}\boldsymbol{Q}\end{bmatrix}}_{(n+d)\times d \text{ matrix}} , \quad \tilde{\boldsymbol{b}} := \underbrace{\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{0}\end{bmatrix}}_{(n+d)\times 1 \text{ vector}} ,$$

  where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ is any orthogonal matrix.
  That is, add d fictitious labeled data $(\sqrt{\lambda}\boldsymbol{q}_j, 0)$ for $j = 1,2,\dots,d$, where $\boldsymbol{q}_1,\boldsymbol{q}_2,\dots,\boldsymbol{q}_d$ is an orthonormal basis.
- Then $\tilde{\boldsymbol{A}}^\top\tilde{\boldsymbol{A}} = \boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I}$, and $\tilde{\boldsymbol{A}}^\top\tilde{\boldsymbol{b}} = \boldsymbol{A}^\top\boldsymbol{b}$.
- So $\hat{\boldsymbol{w}}_{\text{ols}}$ using $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{b}}$ is the same as $\hat{\boldsymbol{w}}_{\text{ridge}}$ using $\boldsymbol{A}$ and $\boldsymbol{b}$.
- For a similar reason, $\hat{\boldsymbol{w}}_{\text{ridge}}$ has a certain Bayesian interpretation.

Ridge regression: key takeaways

- Always well-defined (for $\lambda > 0$)
- Behavior depends on eigenvectors/eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$; the original "coordinate system" is not relevant here.
- Relative to $\hat{\boldsymbol{w}}_{\text{ols}}$ (when it exists): shrinks $\hat{\boldsymbol{w}}_{\text{ols}}$ along eigenvector directions by amount related to eigenvalue and λ.
- Regularization parameter λ 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

Sparsity

Another form of regularization: only consider *sparse* $\boldsymbol{w}$—i.e., $\boldsymbol{w}$ 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 $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ (both in terms of computing variable values and computing inner product).

Sparse regression methods

For any $T \subseteq \{1, 2, \ldots, d\}$, let $\hat{\boldsymbol{w}}_T :=$ 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

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

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

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.)

Sparse regression methods

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

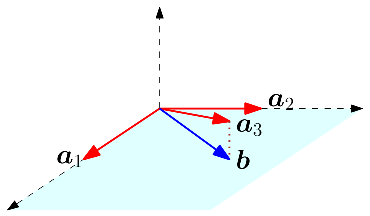
$$\|\boldsymbol{A}\hat{\boldsymbol{w}}_{T \cup \{j\}} - \boldsymbol{b}\|_2^2$$

is minimal, and add this j to T.

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

Gives you a k-sparse solution.

Primarily only effective when columns of $\boldsymbol{A}$ are close to orthogonal.



Aside: l_p norms

For $p \geq 1$,

$$\|\boldsymbol{v}\|_p \;=\; \left(\sum_{j=1}^{d} |v_j|^p\right)^{1/p} .$$

In particular,

$$\|\boldsymbol{v}\|_1 \;=\; \sum_{j=1}^{d} |v_j| .$$

These are *norms* on $\mathbb{R}^d$, so:

- $\|\boldsymbol{u} - \boldsymbol{v}\|_p$ is a valid *metric* on points in $\mathbb{R}^d$, and
- $\|c\boldsymbol{v}\|_p = |c| \cdot \|\boldsymbol{v}\|_p$ for any $\boldsymbol{v} \in \mathbb{R}^d$ and $c \in \mathbb{R}$.

Lasso (Tibshirani, 1994)

Lasso: least absolute shrinkage and selection operator

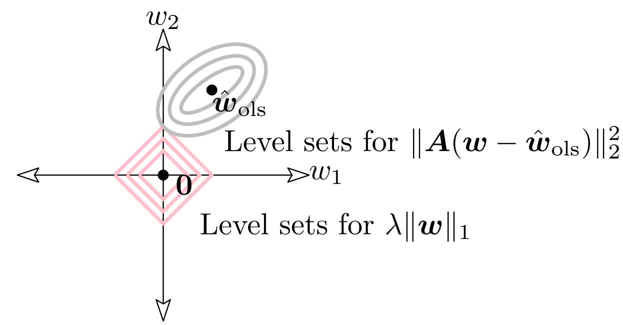
Let $\hat{\boldsymbol{w}}_{\text{lasso}}$ be a minimizer of

$$\boldsymbol{w} \;\mapsto\; \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 .$$

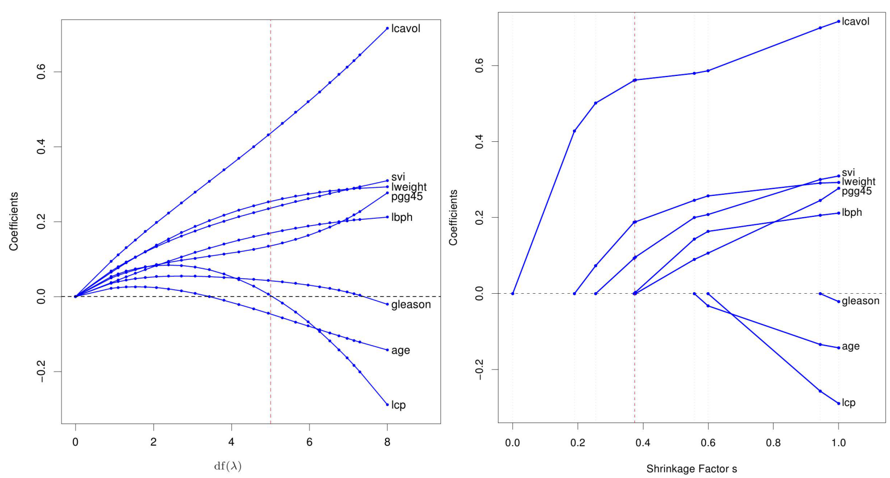
Objective function is convex though not differentiable.

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

$$\|\boldsymbol{A}(\boldsymbol{w} - \hat{\boldsymbol{w}}_{\text{ols}})\|_2^2 + \lambda\|\boldsymbol{w}\|_1 + (\text{stuff not depending on } \boldsymbol{w}) .$$



Coefficient profile (Ridge vs. Lasso)



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

Lasso: theory

Many results, mostly roughly of the following flavor.
Suppose

- $\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{w}_\star, \sigma^2 \boldsymbol{I})$;
- $\boldsymbol{w}_\star$ has $\leq k$ non-zero entries;
- $\boldsymbol{A}$ satisfies some special properties (typically not efficiently checkable);
- $\lambda \gtrsim \sigma\sqrt{2n\log(d)}$;

then

$$\mathbb{E}\left[\|\boldsymbol{w}_\star - \hat{\boldsymbol{w}}_{\text{lasso}}\|_2^2\right] \;\leq\; O\!\left(\frac{\sigma^2 k \log(d)}{n}\right).$$

Closely related to "compressed sensing"; theory involves high-dimensional convex geometry.

Sparse regression: key takeaways

- **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.