ECE595 / STAT598: Machine Learning I
Lecture 02: Regularized Linear Regression

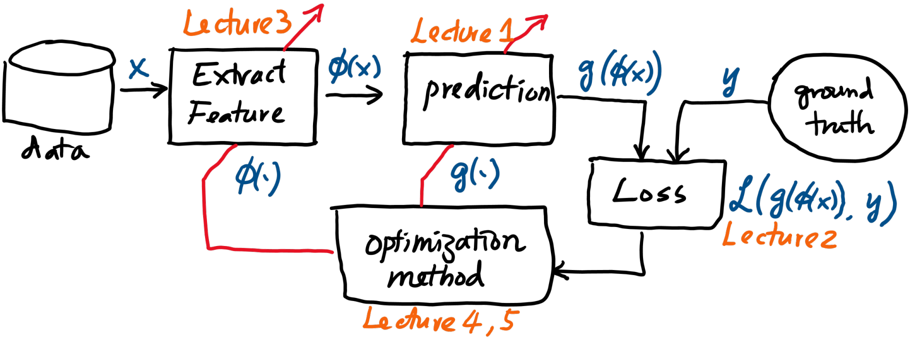
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Outline


Lecture 3
Lecture 1
data
x
Extract Feature
$\phi(x)$
prediction
$g(\phi(x))$
y
ground truth
$\phi(\cdot)$
$g(\cdot)$
Loss
$\mathcal{L}(g(\phi(x)), y)$
Lecture 2
optimization method
Lecture 4, 5

Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 2: Regularization

- Ridge Regression
  - Regularization
  - Parameter
- LASSO Regression
  - Sparsity
  - Algorithm
  - Application

Ridge Regression

- Applies to both over and under determined systems.
- The loss function of the ridge regression is defined as

$$J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

- $\|\boldsymbol{\theta}\|^2$ Regularization function
- λ: Regularization parameter
- The solution of the ridge regression is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}\Big\{\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2\Big\} \\ &= 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}) + 2\lambda\boldsymbol{\theta} = \boldsymbol{0},\end{aligned}$$

  which gives us $\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$.
- Probabilistic interpretation: See Appendix.

Change in Eigen-values

Ridge regression improves the eigen-values:

- Eigen-decomposition of $\boldsymbol{A}^T\boldsymbol{A}$:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T \succeq 0,$$

  where $\boldsymbol{U}$ = eigen-vector matrix, $\boldsymbol{S}$ = eigen-value matrix.

- $\boldsymbol{S}$ is a diagonal matrix with non-negative entries:

$$\boldsymbol{S} = \begin{bmatrix} \clubsuit & & & \\ & \clubsuit & & \\ & & \clubsuit & \\ & & & 0 \end{bmatrix}$$

  See Tutorial on "Linear Algebra".

- Therefore, $\boldsymbol{S} + \lambda\boldsymbol{I}$ is always positive for any $\lambda > 0$, implying that

$$\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I} = \boldsymbol{U}(\boldsymbol{S} + \lambda\boldsymbol{I})\boldsymbol{U}^T \succ 0.$$

Regularization Parameter λ

- The solution of the ridge regression is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

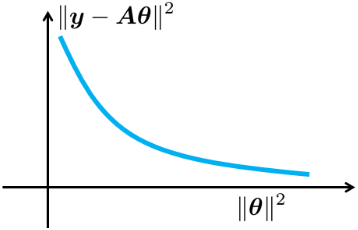
- If $\lambda \to 0$, then $\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$:

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \cancel{\lambda\|\boldsymbol{\theta}\|^2}.$$

- If $\lambda \to \infty$, then $\widehat{\boldsymbol{\theta}} = \boldsymbol{0}$:

$$J(\boldsymbol{\theta}) = \cancel{\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2} + \lambda\|\boldsymbol{\theta}\|^2.$$

- There is a trade-off curve between the two terms by varying λ.



Comparing Vanilla and Ridge

Suppose $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}$ for some ground truth $\boldsymbol{\theta}^*$ and noise vector $\boldsymbol{e}$. Then, the **vanilla linear regression** will give us

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}) \\ &= \boldsymbol{\theta}^* + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{e}\end{aligned}$$

If $\boldsymbol{e}$ has zero mean and variance σ^2, we can show that

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\theta}}] &= \boldsymbol{\theta}^*, \\ \text{Cov}[\widehat{\boldsymbol{\theta}}] &= \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}.\end{aligned}$$

Therefore, the regression coefficients are unbiased but have large variance. We can further show that the mean-squared error (MSE) is

$$\text{MSE}(\widehat{\boldsymbol{\theta}}) = \sigma^2\text{Tr}\left\{(\boldsymbol{A}^T\boldsymbol{A})^{-1}\right\}.$$

Comparing Vanilla and Ridge

On the other hand, if we use ridge regression, then

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(\lambda) &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^* + \boldsymbol{e}) \\ &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^* + (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{e}.\end{aligned}$$

Again, if $\boldsymbol{e}$ is zero mean and has a variance σ^2, then (See Reading List)

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\theta}}(\lambda)] &= (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^* \\ \text{Cov}[\widehat{\boldsymbol{\theta}}(\lambda)] &= \sigma^2(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-T} \\ \text{MSE}[\widehat{\boldsymbol{\theta}}(\lambda)] &= \sigma^2\text{Tr}\{\boldsymbol{W}_\lambda(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{W}_\lambda^T\} + \boldsymbol{\theta}^{*T}(\boldsymbol{W}_\lambda - \boldsymbol{I})^T(\boldsymbol{W}_\lambda - \boldsymbol{I})\boldsymbol{\theta}^*,\end{aligned}$$

where $\boldsymbol{W}_\lambda \overset{\text{def}}{=} (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{A}$. In particular, we can show that

Theorem (Theobald 1974)

For $\lambda < 2\sigma^2\|\boldsymbol{\theta}^*\|^{-2}$, it holds that $MSE(\widehat{\boldsymbol{\theta}}(\lambda)) < MSE(\widehat{\boldsymbol{\theta}})$.

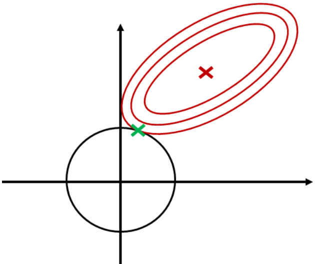
Geometric Interpretation

The following three problems are equivalent

$$
\begin{aligned}
\boldsymbol{\theta}_\lambda^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2 \\
\boldsymbol{\theta}_\alpha^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \qquad \text{subject to} \quad \|\boldsymbol{\theta}\|^2 \le \alpha \\
\boldsymbol{\theta}_\epsilon^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{\theta}\|^2 \qquad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \le \epsilon
\end{aligned}
$$

under an appropriately chosen tuple $(\lambda, \alpha, \epsilon)$.

- Larger λ = Smaller α
- $\boldsymbol{\theta}^*$'s magnitude is tighter bounded



Choosing λ

Because the following three problems are equivalent

$$
\begin{aligned}
\boldsymbol{\theta}_\lambda^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \;\; \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2 \\
\boldsymbol{\theta}_\alpha^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \;\; \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 && \text{subject to} \;\; \|\boldsymbol{\theta}\|^2 \le \alpha \\
\boldsymbol{\theta}_\epsilon^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \;\; \|\boldsymbol{\theta}\|^2 && \text{subject to} \;\; \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \le \epsilon
\end{aligned}
$$

- We can seek λ that satisfies $\|\boldsymbol{\theta}\|^2 \le \alpha$:
  - You know how much $\|\boldsymbol{\theta}\|^2$ would be appropriate.
- We can seek λ that satisfies $\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \le \epsilon$
  - You know how much $\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2$ would be tolerable.
- Other approaches:
  - Akaike's information criterion: Balance model fit with complexity
  - Cross validation: Leave one out
  - Generalized cross-validation: Cross-validation + weight

Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 2: Regularization

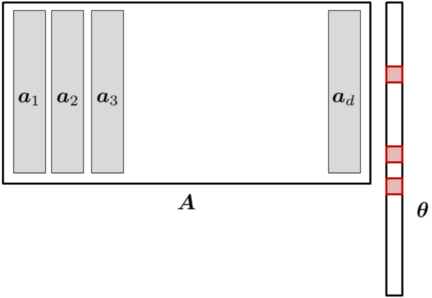
- Ridge Regression
  - Regularization
  - Parameter
- LASSO Regression
  - Sparsity
  - Algorithm
  - Application

LASSO Regression

- An alternative to the Ridge Regression is **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (**LASSO**)
- The loss function is

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

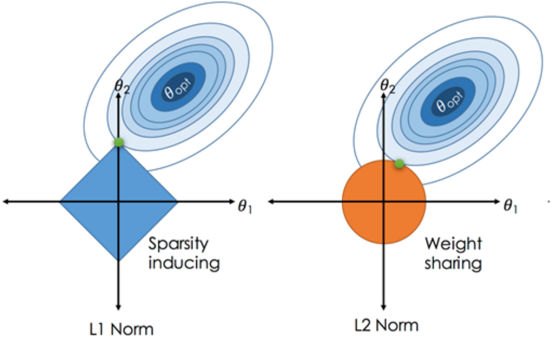
- Intuition behind LASSO: Many features are not active.



Interpreting the LASSO Solution

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

- $\|\boldsymbol{\theta}\|_1$ promotes sparsity of $\boldsymbol{\theta}$. It is the nearest convex approximation to $\|\boldsymbol{\theta}\|_0$, which is the number of non-zeros.
- The difference between ℓ_2 and ℓ_1 [1]:



[1] Figure source: `http://www.ds100.org/`

Why are Sparse Models Useful?

\# non-zeros = 33.51% | 13.58% | 1.21%

- Images are sparse in transform domains, e.g., Fourier and wavelet.
- Intuition: There are more low frequency components and less high frequency components.
- Examples above: $\boldsymbol{A}$ is the wavelet basis matrix. $\boldsymbol{\theta}$ are the wavelet coefficients.
- We can truncate the wavelet coefficients and retain a good image.
- Many image compression schemes are based on this, e.g., JPEG, JPEG2000.

LASSO for Image Reconstruction

Image inpainting via KSVD **dictionary-learning** [2]

- $\boldsymbol{y}$ = image with missing pixels. $\boldsymbol{A}$ = a matrix storing a set of trained feature vectors (called dictionary atoms). $\boldsymbol{\theta}$ = coefficients.
- minimize $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$.
- KSVD = k-means + Singular Value Decomposition (SVD): A method to train the feature vectors that demonstrate sparse representations.

[2] Figure is taken from Mairal, Elad, Sapiro, IEEE T-IP 2008
https://ieeexplore.ieee.org/document/4392496

Shrinkage Operator

The LASSO problem can be solved using a shrinkage operator. Consider a simplified problem (with $\boldsymbol{A} = \boldsymbol{I}$)

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \\
&= \sum_{j=1}^{d} \left\{ \frac{1}{2}(y_j - \theta_j)^2 + \lambda|\theta_j|_1 \right\}
\end{aligned}
$$

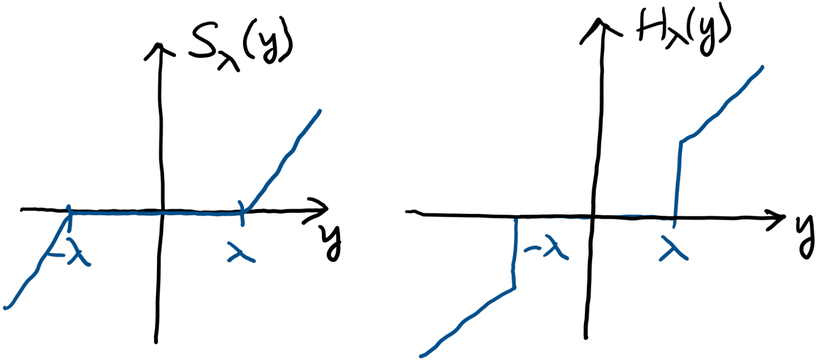
Since the loss is **separable**, the ,optimization is solved when each individual term is minimized. The individual problem

$$
\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmin}} \left\{ \frac{1}{2}(y - \theta)^2 + \lambda|\theta| \right\} \\
&= \max(|y| - \lambda, 0)\text{sign}(y) \\
&\overset{\text{def}}{=} \mathcal{S}_\lambda(y).
\end{aligned}
$$

Proof: See Appendix.

Shrinkage VS Hard Threshold

- The shrinkage operator looks as follows.
- Any number between $[-\lambda, \lambda]$ is "shrink" to zero.
- Try compare with the hard threshold operator $\mathcal{H}_\lambda(y) = y \cdot \mathbf{1}\{|y| \geq \lambda\}$



Algorithms to Solve LASSO Regression

In general, the LASSO problem requires iterative algorithms:

- ISTA Algorithm (Daubechies et al. 2004)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{v}^k = \boldsymbol{\theta}^k - 2\gamma \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^k - \boldsymbol{y})$.
  - $\boldsymbol{\theta}^{k+1} = \max(|\boldsymbol{v}^k| - \lambda, 0)\text{sign}(\boldsymbol{v}^k)$.
- FISTA Algorithm (Beck-Teboulle 2008)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{v}^k = \boldsymbol{\theta}^k - 2\gamma \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta}^k - \boldsymbol{y})$.
  - $\boldsymbol{z}^k = \max(|\boldsymbol{v}^k| - \lambda, 0)\text{sign}(\boldsymbol{v}^k)$.
  - $\boldsymbol{\theta}^{k+1} = \alpha_k \boldsymbol{\theta}^k + (1 - \alpha_k)\boldsymbol{z}^k$.
- ADMM Algorithm (Eckstein-Bertsekas 1992, Boyd et al. 2011)
  - For $k = 1, 2, \ldots$
  - $\boldsymbol{\theta}^{k+1} = (\boldsymbol{A}^T\boldsymbol{A} + \rho \boldsymbol{I})^{-1}(\boldsymbol{A}^T\boldsymbol{y} + \rho \boldsymbol{z}^k - \boldsymbol{u}^k)$
  - $\boldsymbol{z}^{k+1} = \max(|\boldsymbol{\theta}^{k+1} + \boldsymbol{u}^k/\rho| - \lambda/\rho, 0)\text{sign}(\boldsymbol{\theta}^{k+1} + \boldsymbol{u}^k/\rho)$
  - $\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \rho(\boldsymbol{\theta}^{k+1} - \boldsymbol{z}^{k+1})$
- And many others.

Example: Crime Rate Data

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

https://web.stanford.edu/~hastie/StatLearnSparsity/data.html

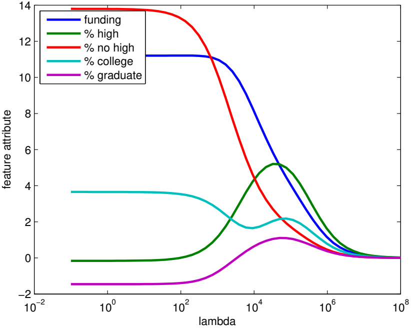
Consider the following two optimizations

$$\widehat{\boldsymbol{\theta}}_1(\lambda) = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad J_1(\boldsymbol{\theta}) \overset{\text{def}}{=} \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

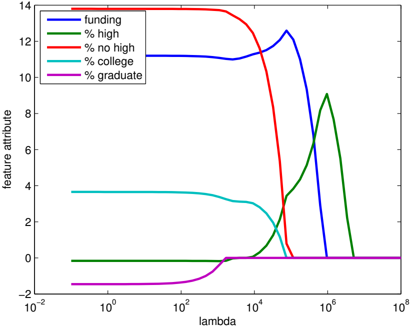
$$\widehat{\boldsymbol{\theta}}_2(\lambda) = \underset{\boldsymbol{\theta}}{\text{argmin}} \quad J_2(\boldsymbol{\theta}) \overset{\text{def}}{=} \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2.$$

Comparison between ℓ-1 and ℓ-2 norm

- Plot $\widehat{\boldsymbol{\theta}}_1(\lambda)$ and $\widehat{\boldsymbol{\theta}}_2(\lambda)$ vs. λ.
- LASSO tells us which factor appears first.
- If we are allowed to use only one feature, then % high is the one.
- Two features, then % high + funding.



Ridge



LASSO

Pros and Cons

Ridge Regression

- (+) Analytic solution, because the loss function is differentiable.
- (+) As such, a lot of well-established theoretical guarantees.
- (+) Algorithm is simple, just one equation.
- (-) Limited interpretability, since the solution is usually a dense vector.
- (-) Does not reflect the nature of certain problems, e.g., sparsity.

LASSO

- (+) Proven applications in many domains, e.g., images and speeches.
- (+) Echoes particularly well in modern deep learning where parameter space is huge.
- (+) Increasing number of theoretical guarantees for special matrices.
- (+) Algorithms are available.
- (-) No closed-form solution. Algorithms are iterative.

Reading List

Ridge Regression

- Stanford CS 229 Note on Linear Algebra
  http://cs229.stanford.edu/section/cs229-linalg.pdf
- Lecture Note on Ridge Regression
  https://arxiv.org/pdf/1509.09169.pdf
- Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society. Series B (Methodological), 36(1), 103-106.

LASSO Regression

- ECE/STAT 695 (Lecture 1)
  https://engineering.purdue.edu/ChanGroup/ECE695.html
- Statistical Learning with Sparsity (Chapter 2)
  https://web.stanford.edu/~hastie/StatLearnSparsity/
- Elements of Statistical Learning (Chapter 3.4)
  https://web.stanford.edu/~hastie/ElemStatLearn/

Appendix

Treating Linear Regression as Maximum-Likelihood

- Minimizing $J(\boldsymbol{\theta})$ is the same as solving a **maximum-likelihood**:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \sum_{n=1}^{N} (\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \exp\left\{ -\sum_{n=1}^{N} (\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2 \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2}{2\sigma^2} \right\} \right\}
\end{aligned}$$

- Assume noise is i.i.d. Gaussian with variance σ^2.
- See Tutorial on Probability

Likelihood Function

- **Likelihood**:

$$p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\boldsymbol{\theta}) = \text{probability density of } \boldsymbol{x} \text{ given } \boldsymbol{\theta}$$

- **Prior**:

$$p_{\Theta}(\boldsymbol{\theta}) = \text{probability density of } \boldsymbol{\theta}$$

- **Posterior**:

$$p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \text{probability density of } \boldsymbol{\theta} \text{ given } \boldsymbol{x}$$

- Bayes Theorem

$$\begin{aligned} p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{p_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})}{\int p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\boldsymbol{\theta})p_{\Theta}(\boldsymbol{\theta})d\boldsymbol{\theta}} \end{aligned}$$

Treating Linear Regression as Maximum-a-Posteriori

- We can modify the MLE by adding a prior

$$p_{\Theta}(\boldsymbol{\theta}) = \exp\left\{ -\frac{\rho(\boldsymbol{\theta})}{\beta} \right\}.$$

- Then, we have a MAP problem:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2}{2\sigma^2} \right\} \right\} \exp\left\{ -\frac{\rho(\boldsymbol{\theta})}{\beta} \right\} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2 + \frac{1}{\beta}\rho(\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\rho(\boldsymbol{\theta}), \qquad \text{where} \quad \lambda = 2\sigma^2/\beta.
\end{aligned}$$

- $\rho(\cdot)$ is called **regularization function**.

Ridge Regression interpreted via a Gaussian prior

- One option: Choose a Gaussian prior

$$\exp\left\{ -\frac{\rho(\boldsymbol{\theta})}{\beta} \right\} = \exp\left\{ -\frac{\|\boldsymbol{\theta}\|^2}{2\sigma_0^2} \right\}$$

- Then, the MAP becomes

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\text{argmax}} \; \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2}{2\sigma^2} \right\} \right\} \exp\left\{ -\frac{\|\boldsymbol{\theta}\|^2}{2\sigma_0^2} \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \; \sum_{n=1}^{N} (\boldsymbol{\theta}^T \boldsymbol{x}^n - y^n)^2 + \underbrace{\frac{\sigma^2}{\sigma_0^2}}_{=\lambda} \|\boldsymbol{\theta}\|^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \; \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\theta}\|^2
\end{aligned}$$

- This is exactly the **ridge regression**.

Proof of the Shrinkage Operator

Let $J(\theta) = \frac{1}{2}(\theta - y)^2 + \lambda|\theta|$.

$$0 = \frac{d}{d\theta}J(\theta) = (\theta - y) + \lambda \text{sign}(\theta).$$

- If $\theta > 0$, then $\theta = y - \lambda$. But since $\theta > 0$, it holds that $y > \lambda > 0$.
- If $\theta < 0$, then $\theta = y + \lambda$. But since $\theta < 0$, it holds that $y < -\lambda < 0$.
- If $\theta = 0$, then $\theta = y$. But since $\theta = 0$, it holds that $y = 0$.
- So the solution is

$$\widehat{\theta} = \begin{cases} y - \lambda, & \text{if} \quad y > 0, \\ 0 & \text{if} \quad y = 0, \\ y + \lambda, & \text{if} \quad y < 0. \end{cases}$$

- This is the same as

$$\widehat{\theta} = \max(|y| - \lambda, 0)\text{sign}(y).$$