Introduction to Machine Learning

Lecture 4 Linear Regression - Bias, Variance, and Regularization

Hongteng Xu

中国人民大学 RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院 Gaoling School of Artificial Intelligence

Outline

Review

- **Polynomial regression:** Formulation, rationality, learning
- **The devil is in the details:** Data preprocessing, stability of learning methods and models, and evaluation (loss function design)
- **Model selection:** AIC and BIC

Outline

Review

- **Polynomial regression:** Formulation, rationality, learning
- **The devil is in the details:** Data preprocessing, stability of learning methods and models, and evaluation (loss function design)
- **Model selection:** AIC and BIC

Today

- Generalized linear regression
- Bias v.s. variance (underfitting, overfitting, ...)
- Regularization methods

Revisit Polynomial Regression

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\| \tag{1}$$

where the Vandermonde matrix $\boldsymbol{X} \in \mathbb{R}^{N\times D}$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{D-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^{D-1} \end{bmatrix} \tag{2}$$

Revisit Polynomial Regression

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\| \tag{1}$$

where the Vandermonde matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{D-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^{D-1} \end{bmatrix} \tag{2}$$

- The polynomial function works as a feature extractor / data representer, mapping each scalar to a D-dimensional feature vector.

Ordinary Linear Regression

- Given arbitrary N D-dimensional features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and their labels $\boldsymbol{y} \in \mathbb{R}^{N}$, an ordinary linear regression is

$$\min_{\boldsymbol{w}} \operatorname{loss}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}) \tag{3}$$

Ordinary Linear Regression

- Given arbitrary N D-dimensional features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and their labels $\boldsymbol{y} \in \mathbb{R}^N$, an ordinary linear regression is

$$\min_{\boldsymbol{w}} \operatorname{loss}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}) \tag{3}$$

- The design of the loss depends on the noise model

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon \tag{4}$$

Ordinary Linear Regression

- Given arbitrary N D-dimensional features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and their labels $\boldsymbol{y} \in \mathbb{R}^{N}$, an ordinary linear regression is

$$\min_{\boldsymbol{w}} \operatorname{loss}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}) \tag{3}$$

- The design of the loss depends on the noise model

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon \tag{4}$$

- Essentially, the learning task is maximizing $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (MLE).

Ordinary Linear Regression

- Given arbitrary N D-dimensional features $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and their labels $\boldsymbol{y} \in \mathbb{R}^{N}$, an ordinary linear regression is

$$\min_{\boldsymbol{w}} \operatorname{loss}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}) \tag{3}$$

- The design of the loss depends on the noise model

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon \tag{4}$$

- Essentially, the learning task is maximizing $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (MLE).
- X are random variables, and a linear regression is interested in the expected value of Y conditioned on X based on a linear predictor, i.e., $\mathbb{E}[Y|X]$.

Random Variables and Instances/Samples

- X: a random variable (r.v.) yielding a distribution P_X, where the random variable can be defined in a D-dimensional space $\mathcal{X}$.

Random Variables and Instances/Samples

- X: a random variable (r.v.) yielding a distribution P_X, where the random variable can be defined in a D-dimensional space $\mathcal{X}$.
- Accordingly, P_X is defined on the space $\mathcal{X}$, and sometimes is denoted as $P_{\mathcal{X}}$.

Random Variables and Instances/Samples

- X: a random variable (r.v.) yielding a distribution P_X, where the random variable can be defined in a D-dimensional space $\mathcal{X}$.
- Accordingly, P_X is defined on the space $\mathcal{X}$, and sometimes is denoted as $P_{\mathcal{X}}$.
- $\boldsymbol{x} \in \mathcal{X}$ is a sample in the space, which can be treated as an instance of the r.v. X, i.e., $\boldsymbol{x} \sim P_X$.

Random Variables and Instances/Samples

- X: a random variable (r.v.) yielding a distribution P_X, where the random variable can be defined in a D-dimensional space $\mathcal{X}$.
- Accordingly, P_X is defined on the space $\mathcal{X}$, and sometimes is denoted as $P_{\mathcal{X}}$.
- $\boldsymbol{x} \in \mathcal{X}$ is a sample in the space, which can be treated as an instance of the r.v. X, i.e., $\boldsymbol{x} \sim P_X$.
- Multiple instances $\boldsymbol{x}$'s often lead to a matrix $\boldsymbol{X}$, and similarly, we denote $\boldsymbol{X} \sim P_X$.

Random Variables and Instances/Samples

- X: a random variable (r.v.) yielding a distribution P_X, where the random variable can be defined in a D-dimensional space $\mathcal{X}$.
- Accordingly, P_X is defined on the space $\mathcal{X}$, and sometimes is denoted as $P_{\mathcal{X}}$.
- $\boldsymbol{x} \in \mathcal{X}$ is a sample in the space, which can be treated as an instance of the r.v. X, i.e., $\boldsymbol{x} \sim P_X$.
- Multiple instances $\boldsymbol{x}$'s often lead to a matrix $\boldsymbol{X}$, and similarly, we denote $\boldsymbol{X} \sim P_X$.
- $\mathbb{E}_{P_X}[X]$ and $\mathbb{V}_{P_X}[X]$ are expectation and variance of the r.v. X.

From Ordinary LR to Generalized Linear Model (GLM)

- GLM is a natural extension of ordinary linear regression, which consists of
  1. An **exponential family** of probability distributions to generate the output.
  2. A **linear predictor** $\eta = X\beta$
  3. A **link function** g: $\mathbb{E}[Y|X] = \mu = g^{-1}(\eta)$.

From Ordinary LR to Generalized Linear Model (GLM)

- GLM is a natural extension of ordinary linear regression, which consists of
  1. An **exponential family** of probability distributions to generate the output.
  2. A **linear predictor** $\eta = X\beta$
  3. A **link function** g: $\mathbb{E}[Y|X] = \mu = g^{-1}(\eta)$.
- The predictor merging input information is linear.
- The link function connecting the prediction and the conditional expectation can be nonlinear (That is why the model is called GLM).

Exponential Family of Probability Distributions

- A parametric distribution $p_X(\boldsymbol{x}|\boldsymbol{\theta})$ having the following form:

$$p_X(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})). \tag{5}$$

- $\boldsymbol{T}(\boldsymbol{x}) : \mathbb{R}^D \mapsto \mathbb{R}^S$: **Sufficient statistic** of the distribution, a function of the data holding all information of the data.

$$I(\boldsymbol{\theta}; \boldsymbol{T}(\boldsymbol{x})) = I(\boldsymbol{\theta}; \boldsymbol{x}) \tag{6}$$

- For **Likelihood ratio:**

$$\frac{p_X(\boldsymbol{x}|\boldsymbol{\theta}_1)}{p_X(\boldsymbol{x}|\boldsymbol{\theta}_2)} = \frac{p_X(\boldsymbol{y}|\boldsymbol{\theta}_1)}{p_X(\boldsymbol{y}|\boldsymbol{\theta}_2)} \quad \text{if} \quad \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y}) \tag{7}$$

- $S = \dim(\boldsymbol{T}(\boldsymbol{x})) = \dim(\boldsymbol{\theta})$.

Exponential Family of Probability Distributions

- A parametric distribution $p_X(\boldsymbol{x}|\boldsymbol{\theta})$ having the following form:

$$p_X(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})). \tag{8}$$

- $\boldsymbol{\eta}(\boldsymbol{\theta}) : \mathbb{R}^S \mapsto \mathbb{R}^S$ is **natural parameter**.
- The natural parameter space, $\{\boldsymbol{\eta} | p_X(\boldsymbol{x}|\boldsymbol{\theta}) \leq \infty\}$, is a convex set.

Exponential Family of Probability Distributions

- A parametric distribution $p_X(\boldsymbol{x}|\boldsymbol{\theta})$ having the following form:

$$p_X(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})). \tag{8}$$

- $\boldsymbol{\eta}(\boldsymbol{\theta}) : \mathbb{R}^S \mapsto \mathbb{R}^S$ is **natural parameter**.
- The natural parameter space, $\{\eta | p_X(\boldsymbol{x}|\boldsymbol{\theta}) \leq \infty\}$, is a convex set.
- $A(\boldsymbol{\theta}) : \mathbb{R}^S \mapsto \mathbb{R}$ is called the **log-partition function** because it is the logarithm of a normalization factor

$$A(\boldsymbol{\theta}) = \log \left(\int_{\boldsymbol{x} \in \mathcal{X}} h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle) \mathrm{d}\boldsymbol{x} \right) \tag{9}$$

- The moments (including mean and variance) of $\boldsymbol{T}(\boldsymbol{x})$ can be derived simply by differentiating $A(\boldsymbol{\theta})$.

Exponential Family of Probability Distributions

- A parametric distribution $p_X(\boldsymbol{x}|\boldsymbol{\theta})$ having the following form:

$$p_X(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})). \tag{8}$$

- $\boldsymbol{\eta}(\boldsymbol{\theta}) : \mathbb{R}^S \mapsto \mathbb{R}^S$ is **natural parameter**.
- The natural parameter space, $\{\boldsymbol{\eta} | p_X(\boldsymbol{x}|\boldsymbol{\theta}) \leq \infty\}$, is a convex set.
- $A(\boldsymbol{\theta}) : \mathbb{R}^S \mapsto \mathbb{R}$ is called the **log-partition function** because it is the logarithm of a normalization factor

$$A(\boldsymbol{\theta}) = \log \left(\int_{\boldsymbol{x} \in \mathcal{X}} h(\boldsymbol{x}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}) \rangle) \mathrm{d}\boldsymbol{x} \right) \tag{9}$$

- The moments (including mean and variance) of $\boldsymbol{T}(\boldsymbol{x})$ can be derived simply by differentiating $A(\boldsymbol{\theta})$.
- $h(\boldsymbol{x}) : \mathbb{R}^D \mapsto \mathbb{R}$ is a non-negative integratable function.

Exponential Family of Probability Distributions

Useful Properties

- For i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$, $\boldsymbol{T}(\boldsymbol{X}) = \sum_{n=1}^{N} \boldsymbol{T}(\boldsymbol{x}_n)$.

Exponential Family of Probability Distributions

Useful Properties

- For i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{T}(\boldsymbol{X}) = \sum_{n=1}^N \boldsymbol{T}(\boldsymbol{x}_n)$.

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\theta}) &= \prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\theta}) \\
&= \prod_{n=1}^N h(\boldsymbol{x}_n) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}_n) \rangle - A(\boldsymbol{\theta}))
\end{aligned}
$$

Exponential Family of Probability Distributions

Useful Properties

- For i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{T}(\boldsymbol{X}) = \sum_{n=1}^N \boldsymbol{T}(\boldsymbol{x}_n)$.

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\theta}) &= \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) \\
&= \prod_{n=1}^{N} h(\boldsymbol{x}_n) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{T}(\boldsymbol{x}_n)\rangle - A(\boldsymbol{\theta})) \\
&= \left(\prod_{n=1}^{N} h(\boldsymbol{x}_n)\right) \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \sum_{n=1}^{N} \boldsymbol{T}(\boldsymbol{x}_n)\rangle - NA(\boldsymbol{\theta})\right)
\end{aligned}
\tag{10}
$$

Exponential Family of Probability Distributions

Useful Properties

- Exponential families have **conjugate priors**
  - In Bayesian probability theory, if the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ is in the same distribution family as the prior distribution $p(\boldsymbol{\theta})$, the prior and posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$.

Exponential Family of Probability Distributions

Useful Properties

- Exponential families have **conjugate priors**
  - In Bayesian probability theory, if the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ is in the same distribution family as the prior distribution $p(\boldsymbol{\theta})$, the prior and posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$.
- The posterior distribution of an exponential-family random variable with a conjugate prior can always be written in closed form. (Important for efficient Bayesian machine learning)

Typical Exponential Families and Their Conjugate Priors

- Normal distribution
- Exponential distribution
- Gamma distribution
- Bernoulli distribution
- Beta distribution
- Poisson distribution
- Categorical distribution
- Geometric distribution
- Multinormal distribution

https://en.wikipedia.org/wiki/Exponential_family

Typical Exponential Families and Their Conjugate Priors

- Normal distribution $\Rightarrow$ Normal/Gamma/Normal-Gamma
- Exponential distribution $\Rightarrow$ Gamma
- Gamma distribution $\Rightarrow$ Gamma
- Bernoulli distribution $\Rightarrow$ Beta
- Poisson distribution $\Rightarrow$ Gamma
- Categorical distribution $\Rightarrow$ Dirichlet
- Geometric distribution $\Rightarrow$ Beta
- Multinormal distribution $\Rightarrow$ Dirichlet

https://en.wikipedia.org/wiki/Conjugate_prior

Revisit Ordinary LR from A Viewpoint of GLM

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon \tag{11}$$

- Exponential family: $y \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{w}, \sigma^2)$
- Linear predictor: $\eta = \boldsymbol{x}^T \boldsymbol{w}$
- Identity link function: $g^{-1}(\eta) = \eta$

Revisit Ordinary LR from A Viewpoint of GLM

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon \tag{11}$$

- Exponential family: $y \sim \mathcal{N}(\boldsymbol{x}^T \boldsymbol{w}, \sigma^2)$
- Linear predictor: $\eta = \boldsymbol{x}^T \boldsymbol{w}$
- Identity link function: $g^{-1}(\eta) = \eta$

The selection of link function is highly relevant to the distribution type: for $y = g^{-1}(\boldsymbol{x}^T \boldsymbol{w}) \sim P$

- Poisson distribution $\Leftrightarrow g(\mu) = \log \mu$
- Gamma distribution $\Leftrightarrow g(\mu) = \frac{1}{\mu}$
- Bernoulli, Categorical, Multinomial $\Leftrightarrow g(\mu) = \log \frac{\mu}{1-\mu}$ (Logit)

https://en.wikipedia.org/wiki/Generalized_linear_model

The Bias of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.

The Bias of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.
- The **bias** of $\hat{\theta}$ relative to θ is

$$\mathrm{Bias}(\hat{\theta}, \theta) = \mathrm{Bias}_\theta[\hat{\theta}] = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta = \mathbb{E}_{x|\theta}[\hat{\theta} - \theta], \tag{12}$$

  which measures the difference between the estimator's expected value and the ground truth.

The Bias of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.
- The **bias** of $\hat{\theta}$ relative to θ is

$$\text{Bias}(\hat{\theta}, \theta) = \text{Bias}_\theta[\hat{\theta}] = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta = \mathbb{E}_{x|\theta}[\hat{\theta} - \theta], \tag{12}$$

  which measures the difference between the estimator's expected value and the ground truth.
- How to understand the notation $\mathbb{E}_{x|\theta}$?

The Bias of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.
- The **bias** of $\hat{\theta}$ relative to θ is

$$\mathrm{Bias}(\hat{\theta}, \theta) = \mathrm{Bias}_\theta[\hat{\theta}] = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta = \mathbb{E}_{x|\theta}[\hat{\theta} - \theta], \tag{12}$$

  which measures the difference between the estimator's expected value and the ground truth.
- How to understand the notation $\mathbb{E}_{x|\theta}$?
- $\mathrm{Bias}(\hat{\theta}, \theta) = 0 \Leftrightarrow$ The estimator $\hat{\theta}$ is unbiased.

Toy Example 1: Is average an unbiased estimation of mean?

Given i.i.d. random variables $\{X_n\}_{n=1}^N$, with expectation μ and variance σ^2.

- Sample average $\hat{\mu} = \frac{1}{N}\sum_{n=1}^N X_n$.

Toy Example 1: Is average an unbiased estimation of mean?

Given i.i.d. random variables $\{X_n\}_{n=1}^N$, with expectation μ and variance σ^2.

- Sample average $\hat{\mu} = \frac{1}{N}\sum_{n=1}^N X_n$.
- We have

$$
\begin{aligned}
\mathrm{Bias}(\hat{\mu}, \mu) =& \mathbb{E}_{X|\mu}[\hat{\mu}] - \mu \\
=& \mathbb{E}_{X|\mu}[\frac{1}{N}\sum_{n=1}^N X_n] - \mu \\
=& \frac{1}{N}\sum_{n=1}^N \mathbb{E}_{X|\mu}[X_n] - \mu \\
=& \frac{N\mu}{N} - \mu = 0
\end{aligned}
\tag{13}
$$

Toy Example 2: What is the unbiased estimation of variance?

- Sample variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$.

Toy Example 2: What is the unbiased estimation of variance?

- Sample variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$.

$$\mathrm{Bias}(\hat{\sigma}^2, \sigma^2) = \mathbb{E}_{X|\sigma^2}[\hat{\sigma}^2] - \sigma^2 = \mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2] - \sigma^2$$

Toy Example 2: What is the unbiased estimation of variance?

- Sample variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$.

$$\mathrm{Bias}(\hat{\sigma}^2, \sigma^2) = \mathbb{E}_{X|\sigma^2}[\hat{\sigma}^2] - \sigma^2 = \mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2] - \sigma^2$$

$$=\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu) - (\hat{\mu} - \mu)]^2] - \sigma^2$$

$$=\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu)^2 - 2(X_n - \mu)(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2]] - \sigma^2$$

Toy Example 2: What is the unbiased estimation of variance?

- Sample variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$.

$$
\begin{aligned}
\mathrm{Bias}(\hat{\sigma}^2, \sigma^2) = \mathbb{E}_{X|\sigma^2}[\hat{\sigma}^2] - \sigma^2 = \mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu) - (\hat{\mu} - \mu)]^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu)^2 - 2(X_n - \mu)(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2]] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2 - \frac{2}{N}(\hat{\mu} - \mu)\sum_{n=1}^{N}(X_n - \mu) + (\hat{\mu} - \mu)^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2 - 2(\hat{\mu} - \mu)^2 + (\hat{\mu} - \mu)^2] - \sigma^2 \\
=&\underbrace{\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2]}_{\sigma^2} - \underbrace{\mathbb{E}_{X|\sigma^2}[(\hat{\mu} - \mu)^2]}_{\frac{1}{N}\sigma^2} - \sigma^2.
\end{aligned}
$$

Toy Example 2: What is the unbiased estimation of variance?

- Sample variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$.

$$
\begin{aligned}
\mathrm{Bias}(\hat{\sigma}^2, \sigma^2) &= \mathbb{E}_{X|\sigma^2}[\hat{\sigma}^2] - \sigma^2 = \mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu) - (\hat{\mu} - \mu)]^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}[(X_n - \mu)^2 - 2(X_n - \mu)(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2]] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2 - \frac{2}{N}(\hat{\mu} - \mu)\sum_{n=1}^{N}(X_n - \mu) + (\hat{\mu} - \mu)^2] - \sigma^2 \\
=&\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2 - 2(\hat{\mu} - \mu)^2 + (\hat{\mu} - \mu)^2] - \sigma^2 \\
=&\underbrace{\mathbb{E}_{X|\sigma^2}[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)^2]}_{\sigma^2} - \underbrace{\mathbb{E}_{X|\sigma^2}[(\hat{\mu} - \mu)^2]}_{\frac{1}{N}\sigma^2} - \sigma^2.
\end{aligned}
\tag{14}
$$

- The unbiased estimation is $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$

The Variance of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.

The Variance of Estimation

- Given a statistical model with parameter θ.
- Given a set of observed data x's and each $x \sim P_\theta(x) = P(x|\theta)$.
- The estimation of θ based on the data points is denoted as $\hat{\theta}$.
- The **variance** of $\hat{\theta}$ is

$$\mathbb{V}[\hat{\theta}] = \mathbb{E}_{x|\theta}[(\hat{\theta} - \mathbb{E}_{x|\theta}[\hat{\theta}])^2] \tag{15}$$

The Trade-off Between Bias and Variance

- Suppose that $\boldsymbol{w}$ is the ground truth parameter of a linear model
- A set of data $(\boldsymbol{X}, \boldsymbol{y})$ are observed and yield

$$y = \underbrace{\boldsymbol{x}^T \boldsymbol{w}}_{f_{\boldsymbol{w}}(\boldsymbol{x})} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

- $\hat{\boldsymbol{w}}$ is the estimator obtained based on the data.

$$\mathrm{MSE} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}}[(y - \hat{y})^2] = \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) + \epsilon - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] = \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2]$$

The Trade-off Between Bias and Variance

- Suppose that $\boldsymbol{w}$ is the ground truth parameter of a linear model
- A set of data $(\boldsymbol{X}, \boldsymbol{y})$ are observed and yield

$$y = \underbrace{\boldsymbol{x}^T \boldsymbol{w}}_{f_{\boldsymbol{w}}(\boldsymbol{x})} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

- $\hat{\boldsymbol{w}}$ is the estimator obtained based on the data.

$$\begin{aligned}
\text{MSE} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}}[(y - \hat{y})^2] = \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) + \epsilon - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] = \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] \\
&= \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] + \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] \\
&= \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])^2] + \underbrace{\mathbb{E}[2(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])(\mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))]}_{=0} \\
&\qquad + \mathbb{E}[(\mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2]
\end{aligned}$$

The Trade-off Between Bias and Variance

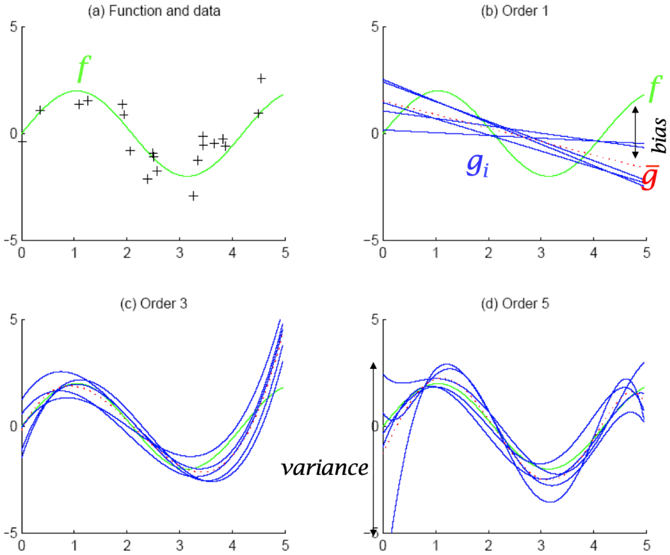
- Suppose that $\boldsymbol{w}$ is the ground truth parameter of a linear model
- A set of data $(\boldsymbol{X}, \boldsymbol{y})$ are observed and yield

$$y = \underbrace{\boldsymbol{x}^T\boldsymbol{w}}_{f_{\boldsymbol{w}}(\boldsymbol{x})} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

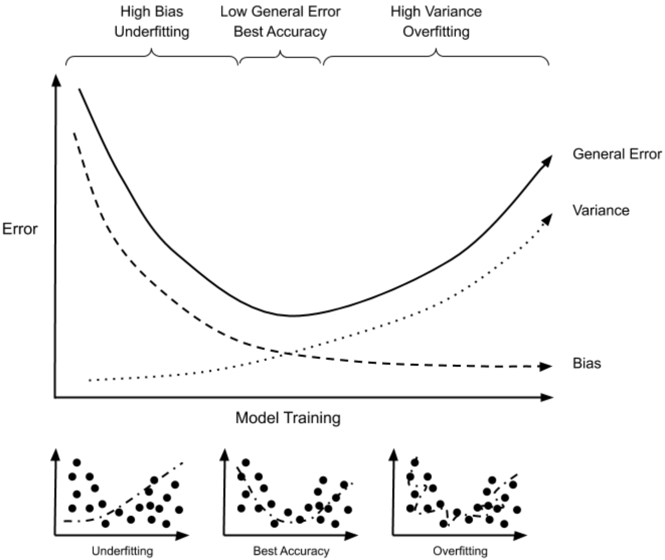
- $\hat{\boldsymbol{w}}$ is the estimator obtained based on the data.

$$\begin{aligned}
\text{MSE} &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}}[(y-\hat{y})^2] = \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) + \epsilon - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] = \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] \\
&= \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] + \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] \\
&= \sigma^2 + \mathbb{E}[(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])^2] + \underbrace{\mathbb{E}[2(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])(\mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))]}_{=0} \\
&\qquad + \mathbb{E}[(\mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})] - f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}))^2] \\
&= \underbrace{\sigma^2}_{\text{Irreducible Noise}} + \underbrace{(f_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])^2}_{\text{Bias}^2(f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}), f_{\boldsymbol{w}}(\boldsymbol{x}))} + \underbrace{\mathbb{E}[(f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}) - \mathbb{E}[f_{\hat{\boldsymbol{w}}}(\boldsymbol{x})])^2]}_{\text{Variance}}
\end{aligned} \tag{17}$$

Bias-Variance Trade-off to Avoid Overfitting and Underfitting



Bias-Variance Trade-off to Avoid Overfitting and Underfitting



Can We Learn Complicated Models from Sparse Data?

- Overfitting: Model complexity $\gg$ data complexity
  - The number of model parameters is larger than that of data points
  - **Case 1:** The model is wrongly complicated $\Rightarrow$ we need to simplify the model
  - **Case 2:** The model is with reasonable complexity but the data are insufficient $\Rightarrow$ more common, and we need to introduce more side information.

Can We Learn Complicated Models from Sparse Data?

- Overfitting: Model complexity $\gg$ data complexity
  - The number of model parameters is larger than that of data points
  - **Case 1:** The model is wrongly complicated $\Rightarrow$ we need to simplify the model
  - **Case 2:** The model is with reasonable complexity but the data are insufficient $\Rightarrow$ more common, and we need to introduce more side information.
- Underfitting: Model complexity $\ll$ data complexity
  - The number of model parameters is smaller than that of data points
- To learn complicated models from sparse data, we need to impose side information on the model parameters (as **regularizers**)

Ridge Regression: MSE with L2 Regularization

- Ridge regression:

$$\min_{\boldsymbol{w}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2}_{L(\boldsymbol{w})} \tag{18}$$

Ridge Regression: MSE with L2 Regularization

- Ridge regression:

$$\min_{\boldsymbol{w}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2}_{L(\boldsymbol{w})} \tag{18}$$

- Consider the data fidelity and penalize the energy of parameters.

Ridge Regression: MSE with L2 Regularization

- Ridge regression:

$$\min_{\boldsymbol{w}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2}_{L(\boldsymbol{w})} \tag{18}$$

- Consider the data fidelity and penalize the energy of parameters.
- Closed form solution:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\lambda\boldsymbol{w} = \boldsymbol{0}$$

Ridge Regression: MSE with L2 Regularization

- Ridge regression:

$$\min_{\boldsymbol{w}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2}_{L(\boldsymbol{w})} \tag{18}$$

- Consider the data fidelity and penalize the energy of parameters.
- Closed form solution:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\lambda\boldsymbol{w} = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{19}$$

The name "ridge" refers to the shape along the diagonal of $\boldsymbol{I}$.

Ridge Regression: MSE with L2 Regularization

- Ridge regression:

$$\min_{\boldsymbol{w}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2}_{L(\boldsymbol{w})} \tag{18}$$

- Consider the data fidelity and penalize the energy of parameters.
- Closed form solution:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\lambda\boldsymbol{w} = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{19}$$

  The name "ridge" refers to the shape along the diagonal of $\boldsymbol{I}$.
- Stochastic gradient descent:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \tau\nabla_{\boldsymbol{w}_t}L \tag{20}$$

Ridge Regression: A Bayesian Viewpoint

- Data model:

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{21}$$

- Model prior:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D \gamma^2) \tag{22}$$

Ridge Regression: A Bayesian Viewpoint

- Data model:

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{21}$$

- Model prior:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D \gamma^2) \tag{22}$$

- Maximizing a posterior (MAP): (**Derive It**)

$$\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) \propto \max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \underbrace{p(\boldsymbol{w}|\boldsymbol{X})}_{p(\boldsymbol{w})}$$

Ridge Regression: A Bayesian Viewpoint

- Data model:

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{21}$$

- Model prior:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D \gamma^2) \tag{22}$$

- Maximizing a posterior (MAP): **(Derive It)**

$$\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) \propto \max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \underbrace{p(\boldsymbol{w}|\boldsymbol{X})}_{p(\boldsymbol{w})} \Rightarrow \max_{\boldsymbol{w}} \prod_n p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) p(\boldsymbol{w})$$

$$\Rightarrow \min_{\boldsymbol{w}} - \sum_n \log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) - \log p(\boldsymbol{w})$$

Ridge Regression: A Bayesian Viewpoint

- Data model:

$$y = \boldsymbol{x}^T \boldsymbol{w} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{21}$$

- Model prior:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D \gamma^2) \tag{22}$$

- Maximizing a posterior (MAP): **(Derive It)**

$$\begin{aligned} &\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) \propto \max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \underbrace{p(\boldsymbol{w}|\boldsymbol{X})}_{p(\boldsymbol{w})} \Rightarrow \max_{\boldsymbol{w}} \prod_n p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) p(\boldsymbol{w}) \\ &\Rightarrow \min_{\boldsymbol{w}} -\sum_n \log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) - \log p(\boldsymbol{w}) \Rightarrow \min_{\boldsymbol{w}} \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{D}{2\gamma^2} \|\boldsymbol{w}\|_2^2 + C. \end{aligned} \tag{23}$$

Some Variants of Ridge Regression

Tikhonov regularization:

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\Gamma}\boldsymbol{w}\|_2^2 \tag{24}$$

- $\boldsymbol{\Gamma}$: Tikhonov matrix
- Derive its closed form solution.

Some Variants of Ridge Regression

Tikhonov regularization:

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{\Gamma}\boldsymbol{w}\|_2^2 \tag{24}$$

- $\boldsymbol{\Gamma}$: Tikhonov matrix
- Derive its closed form solution.

Generalized Tikhonov regularization:

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_{\boldsymbol{P}}^2 + \lambda\|\boldsymbol{w} - \boldsymbol{w}_0\|_{\boldsymbol{Q}}^2 \tag{25}$$

- $\|\boldsymbol{w}\|_{\boldsymbol{A}}^2 = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$
- Derive its closed form solution.

Some Variants of Ridge Regression

Tikhonov regularization:

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{\Gamma}\boldsymbol{w}\|_2^2 \tag{24}$$

- $\boldsymbol{\Gamma}$: Tikhonov matrix
- Derive its closed form solution.

Generalized Tikhonov regularization:

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_{\boldsymbol{P}}^2 + \lambda \|\boldsymbol{w} - \boldsymbol{w}_0\|_{\boldsymbol{Q}}^2 \tag{25}$$

- $\|\boldsymbol{w}\|_{\boldsymbol{A}}^2 = \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}$
- Derive its closed form solution.
- What if $\boldsymbol{P} = \boldsymbol{\Sigma}_y^{-1}$, $\boldsymbol{Q} = \boldsymbol{\Sigma}_w^{-1}$, and $\boldsymbol{w}_0 = \mathbb{E}[\boldsymbol{w}]$?

Lasso: MSE with L1 Regularization

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 \tag{26}$$

- It is also called "Basis pursuit" in the field of signal processing.

Lasso: MSE with L1 Regularization

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{26}$$

- It is also called “Basis pursuit” in the field of signal processing.
- A Bayesian Viewpoint of Lasso: $\boldsymbol{w} \sim \text{Laplace}(0, b\boldsymbol{I}_D)$, so that $p(\boldsymbol{w}) = \frac{1}{(2b)^D}\exp(-\frac{\|\boldsymbol{w}\|_1}{b})$. (Derive the MAP optimization problem)

Lasso: MSE with L1 Regularization

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{26}$$

- It is also called "Basis pursuit" in the field of signal processing.
- A Bayesian Viewpoint of Lasso: $\boldsymbol{w} \sim \mathrm{Laplace}(0, b\boldsymbol{I}_D)$, so that $p(\boldsymbol{w}) = \frac{1}{(2b)^D}\exp(-\frac{\|\boldsymbol{w}\|_1}{b})$. (Derive the MAP optimization problem)
- MAP:

$$\begin{aligned} &\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) \propto \max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \underbrace{p(\boldsymbol{w}|\boldsymbol{X})}_{p(\boldsymbol{w})} \Rightarrow \max_{\boldsymbol{w}} \prod_n p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) p(\boldsymbol{w}) \\ &\Rightarrow \min_{\boldsymbol{w}} -\sum_n \log p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) - \log p(\boldsymbol{w}) \Rightarrow \min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{1}{b}\|\boldsymbol{w}\|_1 + C. \end{aligned} \tag{27}$$

Ridge Regression v.s. Lasso

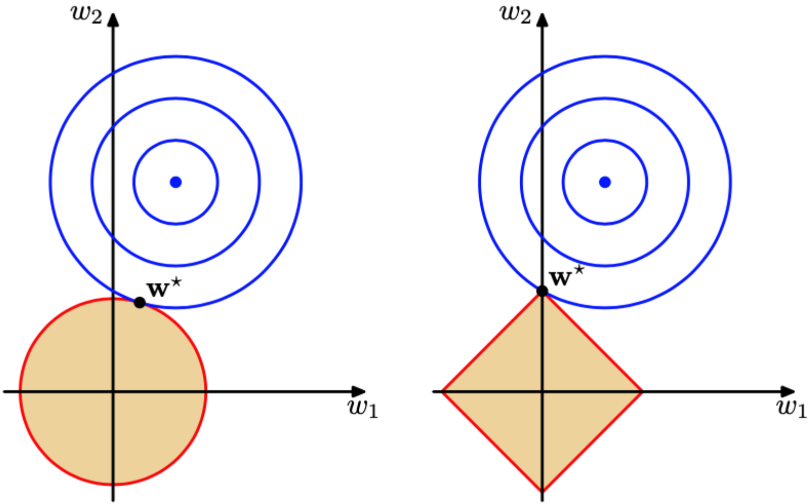
Ridge regression:

- Penalize the energy of parameters.
- Strictly convex and easy to solve with linear convergence.

Lasso:

- Penalize the sparsity of parameters (benefits for model and feature selection).
- Convex but nonsmooth, relatively hard to solve with sublinear convergence.

Ridge Regression v.s. Lasso



Optimization Methods of Lasso Regression

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{28}$$

Soft-thresholding: When $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_D] \in \mathbb{R}^{N\times D}$ are orthonormal ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_D$):

- The solution of ordinary least squares (OLS) is

$$\hat{\boldsymbol{w}}^{(OLS)} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{I}_D\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}.$$

Optimization Methods of Lasso Regression

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{28}$$

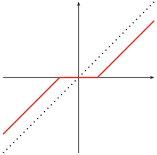
Soft-thresholding: When $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_D] \in \mathbb{R}^{N\times D}$ are orthonormal ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_D$):

- The solution of ordinary least squares (OLS) is

$$\hat{\boldsymbol{w}}^{(OLS)} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{I}_D\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}. \tag{29}$$

- The solution of lasso also has a closed form:

$$\hat{w}_d = S_\lambda(\hat{w}_d^{(OLS)}) = \text{sign}(\hat{w}_d^{(OLS)}) \max\{0, |\hat{w}_d^{(OLS)}| - \lambda\}, \quad \forall d = 1, ..., D. \tag{30}$$



Optimization Methods of Lasso Regression

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1, \quad \text{where } \boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_D] \tag{31}$$

Iterative soft-thresholding for general situations: Although $\boldsymbol{X}^T\boldsymbol{X} \neq \boldsymbol{I}_D$, we can construct orthonormal vectors column-wisely and update parameters iteratively.

- In the t-th iteration, for $d = 1, ..., D$:

$$\begin{aligned}
\hat{w}_d^{(t+1)} &= \arg\min_{w} \|\boldsymbol{y} - \underbrace{\sum_{i\neq d} \boldsymbol{x}_i w_i^{(t)}}_{\boldsymbol{X}_{-d}\boldsymbol{w}_{-d}^{(t)}} - \boldsymbol{x}_d w_d\|_2^2 + \lambda|w| \\
&= \arg\min_{w} \left\| \frac{1}{\|\boldsymbol{x}_d\|_2}(\boldsymbol{y} - \boldsymbol{X}_{-d}\boldsymbol{w}_{-d}^{(t)}) - w \underbrace{\frac{\boldsymbol{x}_d}{\|\boldsymbol{x}_d\|_2}}_{\text{orthonormal}} \right\|_2^2 + \frac{\lambda}{\|\boldsymbol{x}_d\|_2^2}|w| \\
&= S_{\frac{\lambda}{\|\boldsymbol{x}_d\|_2^2}} \left(\frac{\boldsymbol{x}_d^T(\boldsymbol{y} - \boldsymbol{X}_{-d}\boldsymbol{w}_{-d}^{(t)})}{\|\boldsymbol{x}_d\|_2^2} \right)
\end{aligned} \tag{32}$$

Other Methods for Sparse Model Parameters

Lasso

- ADMM (Alternating Direction Method of Multiplier)
- LARS (Least Angle Regression) ...

Other Methods for Sparse Model Parameters

Lasso

- ADMM (Alternating Direction Method of Multiplier)
- LARS (Least Angle Regression) ...

Stronger sparsity: **L0 Regularization and Hard Thresholding**

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0 \tag{33}$$

Other Methods for Sparse Model Parameters

Lasso

- ADMM (Alternating Direction Method of Multiplier)
- LARS (Least Angle Regression) ...

Stronger sparsity: **L0 Regularization and Hard Thresholding**

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0 \tag{33}$$

Weaker sparsity: **Elastic net Regularization**

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|_2^2 \tag{34}$$

Other Methods for Sparse Model Parameters

Lasso

- ADMM (Alternating Direction Method of Multiplier)
- LARS (Least Angle Regression) ...

Stronger sparsity: **L0 Regularization and Hard Thresholding**

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0 \tag{33}$$

Weaker sparsity: **Elastic net Regularization**

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|_2^2 \tag{34}$$

How to interpret it from a Bayesian viewpoint?

Extensions: How to Deal With Outliers?

MAE

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_1 \tag{35}$$

Extensions: How to Deal With Outliers?

MAE

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_1 \tag{35}$$

IRLS (Iteratively Reweighted Least Squares)

$$\begin{aligned}
\boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} \alpha_n(\boldsymbol{w}^{(t)})|y_n - \boldsymbol{x}_n^T\boldsymbol{w}|^2 \\
&= \arg\min_{\boldsymbol{w}} \| \underbrace{\operatorname{diag}^{\frac{1}{2}}(\boldsymbol{\alpha}(\boldsymbol{w}^{(t)}))}_{\boldsymbol{A}^{(t)\,\frac{1}{2}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_2^2
\end{aligned}$$

Extensions: How to Deal With Outliers?

MAE

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_1 \tag{35}$$

IRLS (Iteratively Reweighted Least Squares)

$$\begin{aligned}\boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} \alpha_n(\boldsymbol{w}^{(t)})|y_n - \boldsymbol{x}_n^T\boldsymbol{w}|^2 \\ &= \arg\min_{\boldsymbol{w}} \|\underbrace{\operatorname{diag}^{\frac{1}{2}}(\boldsymbol{\alpha}(\boldsymbol{w}^{(t)}))}_{\boldsymbol{A}^{(t)\frac{1}{2}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_2^2 = (\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{y}\end{aligned}$$

Extensions: How to Deal With Outliers?

MAE

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_1 \tag{35}$$

IRLS (Iteratively Reweighted Least Squares)

$$\begin{aligned}\boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} \alpha_n(\boldsymbol{w}^{(t)})|y_n - \boldsymbol{x}_n^T\boldsymbol{w}|^2 \\ &= \arg\min_{\boldsymbol{w}} \| \underbrace{\operatorname{diag}^{\frac{1}{2}}(\boldsymbol{\alpha}(\boldsymbol{w}^{(t)}))}_{\boldsymbol{A}^{(t)\frac{1}{2}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_2^2 = (\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{y}\end{aligned} \tag{36}$$

$$\alpha_n^{(0)} = 1, \quad \alpha_n^{(t)} = |y_n - \boldsymbol{x}_n^T\boldsymbol{w}^{(t)}|^{-1}$$

Extensions: How to Deal With Outliers?

MAE

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_1 \tag{35}$$

IRLS (Iteratively Reweighted Least Squares)

$$\begin{aligned}\boldsymbol{w}^{(t+1)} &= \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} \alpha_n(\boldsymbol{w}^{(t)})|y_n - \boldsymbol{x}_n^T\boldsymbol{w}|^2 \\ &= \arg\min_{\boldsymbol{w}} \| \underbrace{\operatorname{diag}^{\frac{1}{2}}(\boldsymbol{\alpha}(\boldsymbol{w}^{(t)}))}_{\boldsymbol{A}^{(t)\frac{1}{2}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_2^2 = (\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^{(t)}\boldsymbol{y}\end{aligned} \tag{36}$$

$$\alpha_n^{(0)} = 1, \quad \alpha_n^{(t)} = |y_n - \boldsymbol{x}_n^T\boldsymbol{w}^{(t)}|^{-1}$$

- It works for p-norm with $p \leq 2$, i.e., $\alpha_n^{(t)} = |y_n - \boldsymbol{x}_n^T\boldsymbol{w}^{(t)}|^{p-2}$
- It works as the MLE of GLM, i.e., $y = f_{\boldsymbol{w}}(\boldsymbol{x})$.

In Summary

- Introduce generalized linear regression problem
- Theoretical analysis of linear regression models and some key concepts of statistical machine learning (bias and variance)
- Typical regularization methods and their Bayesian interpretability

Next...

- Non-linear regression
- Duality and kernelization
- Gaussian process