Linear Regression

Yuling Yan

University of Wisconsin–Madison, Fall 2024

From classification to regression

Classification:

- there is a joint distribution of $(X, Y) \sim \rho$ where typically $X \in \mathbb{R}^d$ and $Y \in \{1, \ldots, K\}$ is discrete
- Goal: given input x, find the label y with the highest posterior probability

$$\arg\max_{y \in \{1,\ldots,K\}} \mathbb{P}\big(Y = y | X = x\big)$$

Regression:

- there is a joint distribution of $(X, Y) \sim \rho$ where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$
- Goal: given input x, find a prediction $f(x)$ for Y conditional on $X = x$, that minimizes MSE

$$\mathbb{E}\big[(Y - f(x))^2 | X = x\big]$$

Target of regression problem

Theorem 4.1

For any random variable Z, we have

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}[(Z-c)^2] = \mathbb{E}[Z].$$

Implications for regression problem:

- Conditional on $X = x$, the optimal prediction for Y that minimizes MSE is

$$f^\star(x) = \mathbb{E}[Y|X=x]$$

- Rewrite the model

$$Y = \underbrace{\mathbb{E}[Y|X]}_{\text{regression function}} + \underbrace{Y - \mathbb{E}[Y|X]}_{\text{mean-zero noise}}$$

Regression problem

We will consider the regression problem in a more straightforward way:

$$y = f^\star(\boldsymbol{x}) + \varepsilon$$

- $\boldsymbol{x} \in \mathbb{R}^d$ is the input, $y \in \mathbb{R}$ is the output
- ε is some mean-zero random noise, e.g., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- $f^\star : \mathbb{R}^d \to \mathbb{R}$ is the *unknown* regression function
- Training data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ satisfying

$$y_i = f^\star(\boldsymbol{x}_i) + \varepsilon_i$$

  where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noise with $\mathbb{E}[\varepsilon_i] = 0$, and
  - in some cases, we assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are deterministic (fixed design)
  - sometimes we may assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \rho_X$ (random design)
- Learn the regression function $f^\star$ based on training data

Overview

- **Linear regression:** model the regression function $f^\star$ as a linear function

$$f^\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^\star$$

  where we assume $\boldsymbol{x}$ includes a constant variable 1. Here $\boldsymbol{\beta}^\star \in \mathbb{R}^d$ is the unknown parameter.

- **Nonparametric regression:** assume that

$$f^\star \in \mathcal{F}$$

  where $\mathcal{F}$ is certain function class, e.g.,

  - class of quadratic function
  - class of convex function
  - Reproducing Kernel Hilbert Space (RKHS)

Linear regression: classical setting

Linear regression

- Linear regression:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star + \varepsilon_i \quad (i = 1, \ldots, n)$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are fixed design, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. noise satisfying $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$

- Consider matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$$

Least square estimator

- The most popular estimation method is *least squares*, which estimates $\boldsymbol{\beta}^\star$ by minimizing the residual sum of squares

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

- Ordinary least squares (OLS) estimator:

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

  It has minimizer

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}.$$

- Suppose the noise are i.i.d. Gaussian, then OLS is the MLE

Theoretical properties

- Linear estimator: estimator of the form $\boldsymbol{AY}$ for some matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$
- OLS achieves the minimum variance among all linear unbiased estimators
- Furthermore, when the noise is i.i.d. Gaussian, OLS achieves the minimum variance among all unbiased estimators

Theorem 4.2

- **Gauss-Markov:** *The OLS estimator $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}^\star$, i.e. for any linear and unbiased estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^\star$,*
$$\mathsf{cov}(\widehat{\boldsymbol{\beta}}) \preceq \mathsf{cov}(\widetilde{\boldsymbol{\beta}}).$$
- **Cramér-Rao lower bound:** *when $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, the variance of OLS matches the Cramér-Rao lower bound, i.e. for any unbiased estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^\star$,*
$$\mathsf{cov}(\widehat{\boldsymbol{\beta}}) \preceq \mathsf{cov}(\widetilde{\boldsymbol{\beta}}).$$

Cramér-Rao lower bound

- Consider $X_1, \ldots, X_n$ be i.i.d. samples from a density f_θ
- The unknown parameter $\theta \in \Theta$
- Let $T(X_1, \ldots, X_n)$ be any unbiased estimator for θ
- Under some regularity condition,

$$\mathsf{cov}(T(X_1, \ldots, X_n)) \succeq [I(\theta)]^{-1}$$

where $I(\theta)$ is the **Fisher information matrix**

$$\begin{aligned} I(\theta) &= n\mathbb{E}_{X\sim f_\theta}\left[\nabla_\theta \log f_\theta(X) \left[\nabla_\theta \log f_\theta(X)\right]^\top\right] \\ &= -n\mathbb{E}_{X\sim f_\theta}\left[\nabla_\theta^2 \log f_\theta(X)\right] \end{aligned}$$

Implications

- The OLS estimator is the best one among all unbiased estimator for $\boldsymbol{\beta}^\star$ in terms of minimizing MSE (why?)
- Is it also the best estimator among any estimator for $\boldsymbol{\beta}^\star$, including those biased ones?

Implications

- The OLS estimator is the best one among all unbiased estimator for $\boldsymbol{\beta}^\star$ in terms of minimizing MSE (why?)
- Is it also the best estimator among any estimator for $\boldsymbol{\beta}^\star$, including those biased ones?

  — *No! There are biased estimator which can achieve smaller MSE.*

Implications

- The OLS estimator is the best one among all unbiased estimator for $\boldsymbol{\beta}^\star$ in terms of minimizing MSE (why?)
- Is it also the best estimator among any estimator for $\boldsymbol{\beta}^\star$, including those biased ones?

  — *No! There are biased estimator which can achieve smaller MSE.*

- Examples of biased estimator with smaller MSE:
  - James-Stein estimator
  - Ridge regression

  — *shrinkage estimators*

Shrinkage estimator

Bias-variance tradeoff

- Suppose that the unknown parameter is $\boldsymbol{\beta}^\star \in \mathbb{R}^d$
- For any estimator $\widehat{\boldsymbol{\beta}}$ (more generally, any random vector), the mean squared error (MSE) can be decomposed into

$$\underbrace{\mathbb{E}[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2]}_{=:\text{MSE}} = \underbrace{\|\mathbb{E}[\widehat{\boldsymbol{\beta}}\,] - \boldsymbol{\beta}^\star\|_2^2}_{\text{bias}} + \underbrace{\mathsf{tr}(\mathsf{cov}(\widehat{\boldsymbol{\beta}}))}_{\text{variance}}$$

- For unbiased estimator (e.g., OLS), the bias is zero
- By tolerating a small amount of bias we may be able to achieve a larger reduction in variance, thus achieving smaller MSE

James-Stein estimator

- Consider a Gaussian sequence model,

$$\boldsymbol{Y} = \boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$$

  which is a special linear regression by taking $d = n$ and $\boldsymbol{X} = \boldsymbol{I}_n$

- OLS / MLE: $\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} = \boldsymbol{Y}$
- James-Stein estimator:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{JS}} = \left(1 - \frac{n-2}{\|\boldsymbol{Y}\|_2^2}\right)\boldsymbol{Y}$$

Theorem 4.3

James-Stein estimator has smaller MSE than OLS when $n \geq 3$, i.e.,

$$\mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{\mathsf{JS}}) < \mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}) \quad \textit{for any} \quad \boldsymbol{\beta}^\star$$

Implications

- By shrinking the OLS towards zero, we achieve smaller MSE

  — *inadmissability of OLS (or MLE)*

- It is not even necessary to shrink towards zero: for any fixed $\boldsymbol{c} \in \mathbb{R}^n$,

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} := \boldsymbol{Y} - \frac{p-2}{\|\boldsymbol{Y} - \boldsymbol{c}\|_2^2}(\boldsymbol{Y} - \boldsymbol{c})$$

  also satisfy the same property as Theorem 4.3

- Can be extended to linear regression:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{JS}} = \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} - \frac{(d-2)\widehat{\sigma}^2}{\|\boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}\|_2^2} \boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}.$$

Ridge regression

- Ridge regression: ℓ_2-penalized least squares estimator

$$\widehat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

  where λ is the tuning parameter.

- The ridge regression estimator admits closed-form solution:

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

  It is well defined even when $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible

- As $\lambda \to 0$, ridge regression recovers the OLS
- Interpretation as MAP estimator with a Gaussian prior on $\boldsymbol{\beta}^\star$

MAP estimate

Consider observing X from a density $f_{\theta^\star}$, where $\theta^\star \in \Theta$ is unknown

Frequentist's viewpoint: $\theta^\star$ is fixed (though unknown)

- Likelihood function: $f_\theta(X)$ (a function of $\theta \in \Theta$)
- Estimate $\theta^\star$ by the maximizer of the likelihood function

— *maximum likelihood estimation (MLE)*

Bayesian's viewpoint: θ is also random

- We have a prior distribution $g(\theta)$ over Θ, and conditional on θ, $X \sim f_\theta$
- Posterior probability of θ after observing X:

$$\mathbb{P}(\theta|X) = \frac{g(\theta) f_\theta(X)}{\int_\Theta g(\theta') f_{\theta'}(X) \mathrm{d}\theta'} \propto g(\theta) f_\theta(X)$$

- Estimate θ by the maximizer of the posterior probability

— *maximum a posteriori estimation (MAP)*

Properties of ridge regression

Ridge regression:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}.$$

Theorem 4.4

There exists $\lambda_0 > 0$ such that ridge regression $\widehat{\boldsymbol{\beta}}_\lambda$ achieves smaller MSE than OLS estimate

$$\mathsf{MSE}(\widehat{\boldsymbol{\beta}}_\lambda) < \mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}})$$

for any $\lambda \in (0, \lambda_0]$.

Properties of ridge regression

Ridge regression:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Theorem 4.4

There exists $\lambda_0 > 0$ such that ridge regression $\widehat{\boldsymbol{\beta}}_\lambda$ achieves smaller MSE than OLS estimate

$$\mathsf{MSE}(\widehat{\boldsymbol{\beta}}_\lambda) < \mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}})$$

for any $\lambda \in (0, \lambda_0]$.

- To prove this theorem, we need some tool from linear algebra

Singular Value Decomposition (SVD)

For any rank-r matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, it can be expressed as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

- $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ are orthogonal matrices:

$$\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_r], \qquad \boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_r],$$

  where $\{\boldsymbol{u}_i\}_{i=1}^r$ (resp. $\{\boldsymbol{v}_i\}_{i=1}^r$) are orthonormal vectors in $\mathbb{R}^m$ (resp. $\mathbb{R}^n$)

- $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix

$$\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1, \dots, \sigma_r\}$$

  where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of $\boldsymbol{X}$

More about SVD

For any rank-r matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

- Connection to eigen-decomposition

$$\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{U}_\perp\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}^2 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}_{n-r}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}^\top\\ \boldsymbol{U}_\perp^\top\end{bmatrix}$$

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top = \begin{bmatrix}\boldsymbol{V} & \boldsymbol{V}_\perp\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}^2 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}_{d-r}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}^\top\\ \boldsymbol{V}_\perp^\top\end{bmatrix}$$

  where $\boldsymbol{U}_\perp$ (resp. $\boldsymbol{V}_\perp$) is the orthogonal complement of $\boldsymbol{U}$ (resp. $\boldsymbol{V}$)

- The operator (spectral) norm of $\boldsymbol{X}$

$$\|\boldsymbol{X}\| = \sup_{\|\boldsymbol{a}\|_2=1}\|\boldsymbol{X}\boldsymbol{a}\|_2 = \sigma_1$$

- The Frobenius norm of $\boldsymbol{X}$

$$\|\boldsymbol{X}\|_{\mathrm{F}}^2 = \sum_{i=1}^{r}\sigma_i^2$$

Implications to ridge regression

Suppose that the design matrix $\boldsymbol{X}$ has SVD $\boldsymbol{U\Sigma V}^\top$

- Bias-variance decomposition

$$\mathbb{E}[\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\|_2^2] = \|\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta}^\star\|_2^2 + \mathsf{tr}(\mathsf{cov}(\widehat{\boldsymbol{\beta}}_\lambda))$$

- Bias term

$$\|\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta}^\star\|_2^2 = \sum_{i=1}^d \left(\frac{\lambda \widetilde{\beta}_i}{\lambda + \sigma_i^2}\right)^2 \quad \text{where} \quad \widetilde{\boldsymbol{\beta}} = [\boldsymbol{V}, \boldsymbol{V}_\perp]^\top \boldsymbol{\beta}^\star$$

- Variance term

$$\mathsf{cov}(\widehat{\boldsymbol{\beta}}_\lambda) = \sigma^2 \sum_{i=1}^d \left(\frac{\sigma_i}{\lambda + \sigma_i^2}\right)^2$$

- This allows us to prove Theorem 4.4

Linear regression: high-dimensional setting

What happens in high-dimension?

High-dimensional linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$$

where the dimension d is much larger than the sample size n

- OLS fails because $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible
- In general, it is not possible to say something meaningful about $\boldsymbol{\beta}^\star \in \mathbb{R}^d$ from n samples $\boldsymbol{Y} \in \mathbb{R}^n$ (identifibility issue)

What happens in high-dimension?

High-dimensional linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\varepsilon}$$

where the dimension d is much larger than the sample size n

- OLS fails because $\boldsymbol{X}^{\top}\boldsymbol{X}$ is not invertible
- In general, it is not possible to say something meaningful about $\boldsymbol{\beta}^{\star} \in \mathbb{R}^d$ from n samples $\boldsymbol{Y} \in \mathbb{R}^n$ (identifibility issue)
- A meaningful and workable setup: assume $\boldsymbol{\beta}^{\star}$ is sparse, i.e.,

$$s := \|\boldsymbol{\beta}^{\star}\|_0 \equiv |\{j : \beta_j^{\star} \neq 0\}| \ll d$$

Sparse linear regression

High-dimensional linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$$

where $d \geq n$, but $s = \|\boldsymbol{\beta}^\star\|_0 \ll d$

- **Genomics:** only a small subset of genes is expected to be associated with a particular trait or disease
- **Finance and Economics:** only a small subset of macroeconomic variables or market signals may be relevant to stock returns or economic growth
- $\cdots\cdots$

Insights

- Motivated by ridge regression, we may consider

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0$$

- Issue: computationally hard ($\|\cdot\|_0$ is discontinuous, non-convex...)
- Idea: use $\|\cdot\|_1$ instead
- Insights from compressed sensing (noiseless): under certain conditions (known as restricted isometry property), ℓ_1 minimization problem

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Y}$$

has unique minimizer that coincides with the minimizer to

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Y}.$$

LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) estimates $\boldsymbol{\beta}^\star$ by solving the following convex optimization problem:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where:

- $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$: residual sum of squares (RSS).
- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$: ℓ_1-norm penalty.
- $\lambda > 0$: tuning parameter that controls the trade-off between goodness of fit and sparsity.
- Interpretation as MAP estimator with a Laplace prior on $\boldsymbol{\beta}^\star$
- Questions:
  - How to compute LASSO estimate?
  - What is the statistical properties of LASSO?

How to compute LASSO: proximal gradient method

A more general class of convex optimization

Consider unconstrained convex optimization problem of the form

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x})$$

where

- $f(\boldsymbol{x})$: a differentiable, convex function
- $h(\boldsymbol{x})$: a convex, potentially non-differentiable function (e.g., ℓ_1-norm).
- Example: LASSO can be viewed as taking

$$f(\boldsymbol{x}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \qquad h(\boldsymbol{x}) = \lambda\|\boldsymbol{\beta}\|_1.$$

Issue: gradient descent (GD) does not work (due to non-smoothness)

A Proximal View of Gradient Descent

- To motivate proximal gradient methods, we first revisit gradient descent for $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, where $f(\cdot)$ is convex and smooth
- Gradient descent update: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)$
- This is equivalent to

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}} \Big\{ \underbrace{f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle}_{\text{first-order approximation at } \boldsymbol{x}_t} + \underbrace{\frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2}_{\text{proximal term}} \Big\}$$

- Heuristics: search for $\boldsymbol{x}_{t+1}$ that
  - aim to minimize $f(\cdot)$ (through minimizing first-order approximation)
  - remains close to $\boldsymbol{x}_t$ such that first-order approximation at $\boldsymbol{x}_t$ is valid (enforced by proximal term)
- Benefit: minimizing a quadratic function, admits simple solution (i.e., GD)

Proximal gradient method: algorithm

Consider an iterative algorithm: starting from $\boldsymbol{x}_t$, update

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}} \Big\{ \underbrace{f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle}_{\text{first-order approximation at } \boldsymbol{x}_t} + h(\boldsymbol{x}) + \underbrace{\frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2}_{\text{proximal term}} \Big\}$$

- Define proximal operator

$$\mathsf{prox}_h(\boldsymbol{v}) = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ h(\boldsymbol{x}) + \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{v}\|_2^2 \right\}$$

- If this proximal operator is easy to compute, then we can express

$$\boldsymbol{x}_{t+1} = \mathsf{prox}_{\eta h}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

- alternates between gradient updates on f and proximal minimization on h

Proximal gradient method: properties

Proximal gradient algorithm: for $t = 1, 2, \dots$

$$\boldsymbol{x}_{t+1} = \mathsf{prox}_{\eta h}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

- fast convergence when f is convex and L-smooth: take $\eta = 1/L$,

$$F(\boldsymbol{x}_t) - F^\star \le \frac{L}{2t} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

- exponential convergence when f is μ-strongly convex

$$\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 \le (1 - \mu/L)^t \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

- when $h(\boldsymbol{x}) = 0$ when $\boldsymbol{x} \in \mathcal{A}$ and $h(\boldsymbol{x}) = \infty$ otherwise, this gives the projected gradient descent for $\min_{\boldsymbol{x} \in \mathcal{A}} f(\boldsymbol{x})$:

$$\boldsymbol{x}_{t+1} = \mathcal{P}_{\mathcal{A}}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

- Recommended reading material: Lecture 5 of the course Large-Scale Optimization for Data Science

Application to LASSO

- LASSO:

$$f(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \qquad \text{and} \qquad h(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$$

- The proximal operator admits closed-form expression

$$\mathsf{prox}_h(\boldsymbol{v}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{\frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} = \mathsf{shrink}_\lambda(\boldsymbol{v})$$

where $\mathsf{shrink}_\lambda(\cdot)$ applies entrywise shrinkage to $\boldsymbol{v}$ towards zero:

$$[\mathsf{shrink}_\lambda(\boldsymbol{v})]_j = \begin{cases} v_j - \lambda, & \text{if } v_j \geq \lambda, \\ v_j + \lambda, & \text{if } v_j \leq -\lambda, \\ 0, & \text{otherwise.} \end{cases}$$

- Proximal gradient algorithm for LASSO:

$$\boldsymbol{\beta}_{t+1} = \mathsf{shrink}_{\eta\lambda}\big(\boldsymbol{\beta}_t - 2\eta\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}_t + 2\eta\boldsymbol{X}^\top\boldsymbol{Y}\big)$$

Statistical properties of LASSO

Setup

LASSO:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \Big\{ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \Big\},$$

- Independent, sub-Gaussian noise $\|\varepsilon_i\|_{\psi_2} \le \sigma$
- Sparsity: $n \gg s \log d$
- Theory-informed tuning parameter selection:

$$\lambda \asymp \sigma\sqrt{n \log d}$$

- Question:
    - Does LASSO recover the support of $\boldsymbol{\beta}^\star$?
    - Does LASSO provide reliable estimate for $\boldsymbol{\beta}^\star$?

Optimality condition

The optimality condition for unconstrained convex optimization

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$$

- if f is smooth: $\nabla f(\widehat{\boldsymbol{x}}) = \mathbf{0}$
- in general (when f might not be smooth): $\mathbf{0} \in \partial f(\widehat{\boldsymbol{x}})$

Here $\partial f(\boldsymbol{x}) \subseteq \mathbb{R}^d$ is the **subgradient** of the confex function f at $\boldsymbol{x}$:

$$\boldsymbol{g} \in \partial f(\boldsymbol{x}) \quad \iff \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top (\boldsymbol{y} - \boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{y} \in \mathbb{R}^d$$

Check (in homework):

- if f is smooth at $\boldsymbol{x}$: $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$
- the optimality condition for LASSO is: for each $1 \leq j \leq d$

$$\left[\boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})\right]_j \quad \begin{cases} = \lambda \cdot \mathsf{sign}(\widehat{\beta}_j) & \text{if} \quad \widehat{\beta}_j \neq 0 \\ \in [-\lambda, \lambda] & \text{if} \quad \widehat{\beta}_j = 0 \end{cases}$$

Model selection consistency

- Let $S = \{j : \beta_j^\star \neq 0\}$ be the support set (nonzero coefficients) and S^c be its complement.
- **Irrepresentable condition:**

$$\|\boldsymbol{X}_{S^c}^\top \boldsymbol{X}_S (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1} \boldsymbol{\beta}_S^\star\|_\infty < 1,$$

  where $\boldsymbol{X}_S$ and $\boldsymbol{X}_{S^c}$ as submatrices of $\boldsymbol{X}$ with columns corresponding to S and S^c, and $\boldsymbol{\beta}_S^\star$ is the sub-vector of $\boldsymbol{\beta}^\star$ corresponding to S

- **Model Selection Consistency:** If the irrepresentable condition holds, under certain assumptions, the Lasso estimator satisfies:

$$\mathbb{P}(\widehat{S} = S) \to 1 \quad \text{as } n \to \infty,$$

  where $\widehat{S} = \{j : \widehat{\beta}_j \neq 0\}$.

Estimation guarantees

- **Restricted eigenvalue condition:** For any $\boldsymbol{v} \in \mathbb{R}^p$ such that $\|\boldsymbol{v}_{S^c}\|_1 \leq 3\|\boldsymbol{v}_S\|_1$, the restricted eigenvalue condition is:

$$\min_{\|\boldsymbol{v}\|_2=1, \|\boldsymbol{v}_{S^c}\|_1 \leq 3\|\boldsymbol{v}_S\|_1} \boldsymbol{v}^\top \Big(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\Big)\boldsymbol{v} > 0.$$

  This is satisfied by e.g., i.i.d. Gaussian matrix $\boldsymbol{X}$.

- **Estimation error:** If the restricted eigenvalue condition holds, under certain assumptions, the LASSO estimator satisfies:

$$\frac{1}{n}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \lesssim \sigma^2 s \frac{\log d}{n},$$

  and

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \lesssim \sigma s \sqrt{\frac{\log d}{n}}.$$

Reference

Model selection:

- *Peng Zhao, and Bin Yu. "On model selection consistency of Lasso." The Journal of Machine Learning Research 7 (2006): 2541-2563.*
- *Martin J. Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)." IEEE transactions on information theory 55.5 (2009): 2183-2202.*

Estimation error bounds:

- *Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector." Annals of Statistics 37.4 (2009): 1705-1732.*