Sparsity in Machine Learning

EN.580.709 - Fall 2019

Supervised Learning

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)
- We are given a sample S of n pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled i.i.d from $\mathcal{D}$ defined over $\mathcal{X} \times \mathcal{Y}$

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)
- We are given a sample S of n pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled i.i.d from $\mathcal{D}$ defined over $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$: set of functions $h : \mathcal{X} \to \mathcal{Y}$

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)
- We are given a sample S of n pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled i.i.d from $\mathcal{D}$ defined over $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$: set of functions $h : \mathcal{X} \to \mathcal{Y}$
- L: Loss Function $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)
- We are given a sample S of n pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled i.i.d from $\mathcal{D}$ defined over $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$: set of functions $h : \mathcal{X} \to \mathcal{Y}$
- L: Loss Function $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

Goal: Use the labeled sample S to select a function $h \in \mathcal{H}$ that has small generalization error (or risk)

$$R(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} L(h(\mathbf{x}), y)$$

Supervised Learning

- $\mathcal{X}$: input space, or set of all possible examples (e.g. $\mathcal{X} \subseteq \mathbb{R}^m$)
- $\mathcal{Y}$: label space, or set of all possible labels or target values; (e.g., $\mathcal{Y} = \{0, 1\}$)
- We are given a sample S of n pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled i.i.d from $\mathcal{D}$ defined over $\mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$: set of functions $h : \mathcal{X} \to \mathcal{Y}$
- L: Loss Function $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

Goal: Use the labeled sample S to select a function $h \in \mathcal{H}$ that has small generalization error (or risk)

$$R(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} L(h(\mathbf{x}), y)$$

We only get to measure the *empirical risk*,

$$\hat{R}_S(h) = \frac{1}{n} \sum_i L(h(\mathbf{x}_i), y_i)$$

Linear Regression

- $\mathbf{x}_i \in \mathbb{R}^m$ feature vector of *predictors* $(1, x_2, \ldots, x_m)$
- $y_i \in R$

Linear Regression

- $\mathbf{x}_i \in \mathbb{R}^m$ feature vector of *predictors* $(1, x_2, \ldots, x_m)$
- $y_i \in R$
- Hypothesis as linear combination of predictors:

$$h(\mathbf{x}_i) = \beta_0 + \sum_{j=2}^{m} \beta_i x_{i,j} = \boldsymbol{\beta}^T \mathbf{x}$$

Linear Regression

- $\mathbf{x}_i \in \mathbb{R}^m$ feature vector of *predictors* $(1, x_2, \ldots, x_m)$
- $y_i \in R$
- Hypothesis as linear combination of predictors:

$$h(\mathbf{x}_i) = \beta_0 + \sum_{j=2}^{m} \beta_i x_{i,j} = \boldsymbol{\beta}^T \mathbf{x}$$

We follow an empirical risk minimization approach, let $L(y_i, y_j) = (y_i - y_j)^2$,

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)$$

Linear Regression

- $\mathbf{x}_i \in \mathbb{R}^m$ feature vector of *predictors* $(1, x_2, \ldots, x_m)$
- $y_i \in R$
- Hypothesis as linear combination of predictors:

$$h(\mathbf{x}_i) = \beta_0 + \sum_{j=2}^{m} \beta_i x_{i,j} = \boldsymbol{\beta}^T \mathbf{x}$$

We follow an empirical risk minimization approach, let $L(y_i, y_j) = (y_i - y_j)^2$,

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n^T \end{bmatrix} \in \mathbb{R}^m$$

Linear Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Linear Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- Typically $n \geq m$ (and $\mathbf{X}$ full row rank)

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \mathbf{0} \;\Rightarrow\; \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

Linear Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- Typically $n \geq m$ (and $\mathbf{X}$ full row rank)

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \mathbf{0} \;\Rightarrow\; \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

- If $n < m$, infinite solutions. One possibility is the one with minimal ℓ_2 norm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 \;\text{ s.t. }\; \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

Linear Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- Typically $n \geq m$ (and $\mathbf{X}$ full row rank)

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \mathbf{0} \;\Rightarrow\; \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

- If $n < m$, infinite solutions. One possibility is the one with minimal ℓ_2 norm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 \;\text{ s.t. }\; \mathbf{X}\boldsymbol{\beta} = \mathbf{y} \quad \Rightarrow \hat{\boldsymbol{\beta}} = \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{y}$$

Linear Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

- Typically $n \geq m$ (and $\mathbf{X}$ full row rank)

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \mathbf{0} \;\Rightarrow\; \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- If $n < m$, infinite solutions. One possibility is the one with minimal ℓ_2 norm:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 \;\text{ s.t. }\; \mathbf{X}\boldsymbol{\beta} = \mathbf{y} \quad \Rightarrow \hat{\boldsymbol{\beta}} = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}$$

Ridge (Regularized) Regression

$$\min_{\boldsymbol{\beta}} \mathcal{L} = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^T\mathbf{y}\right)$$

Polynomial Regression

Let $x \in \mathbb{R}$,

$$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m = \boldsymbol{\beta}^T \phi(x)$$

Polynomial Regression

Let $x \in \mathbb{R}$,

$$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m = \boldsymbol{\beta}^T \phi(x)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1, & \phi_1(x_1), & \ldots, & \phi_m(x_1) \\ 1, & \phi_1(x_2), & \ldots, & \phi_m(x_2) \\ & & \vdots & \\ 1, & \phi_1(x_n), & \ldots, & \phi_m(x_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

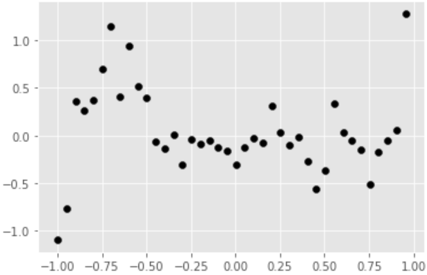
Polynomial Regression

Let $x \in \mathbb{R}$,

$$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m = \boldsymbol{\beta}^T \phi(x)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1, & \phi_1(x_1), & \ldots, & \phi_m(x_1) \\ 1, & \phi_1(x_2), & \ldots, & \phi_m(x_2) \\ & & \vdots & \\ 1, & \phi_1(x_n), & \ldots, & \phi_m(x_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

$$y_i = h^*(x_i) + v_i, \quad v_i \sim \mathcal{N}(0, \sigma^2)$$



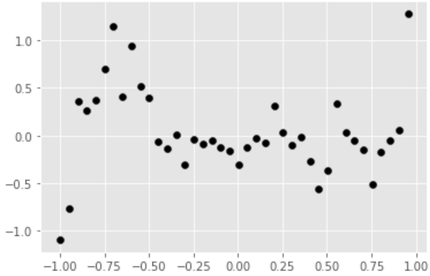
Polynomial Regression

Let $x \in \mathbb{R}$,

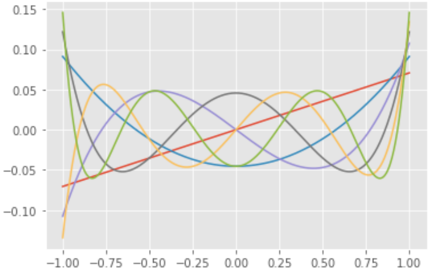
$$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m = \boldsymbol{\beta}^T \phi(x)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1, & \phi_1(x_1), & \ldots, & \phi_m(x_1) \\ 1, & \phi_1(x_2), & \ldots, & \phi_m(x_2) \\ & & \vdots & \\ 1, & \phi_1(x_n), & \ldots, & \phi_m(x_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

$y_i = h^*(x_i) + v_i, \quad v_i \sim \mathcal{N}(0, \sigma^2)$



$\phi_j(x), \ 1 \le j \le m$

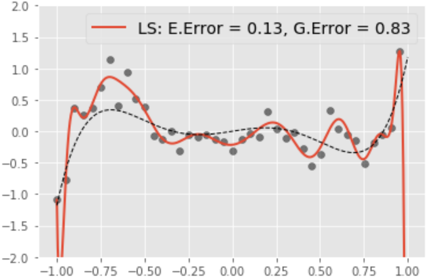


Polynomial Regression $(n > m)$

Linear Regression
$\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

Polynomial Regression $(n > m)$

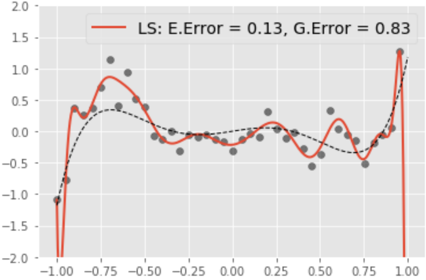
Linear Regression
$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

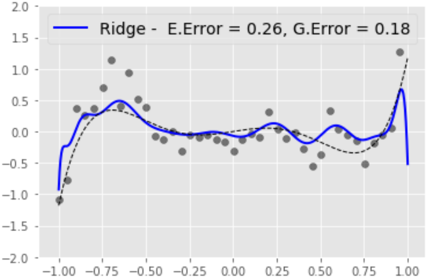


Polynomial Regression $(n > m)$

Linear Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Ridge Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$



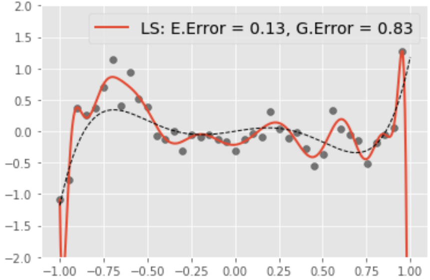


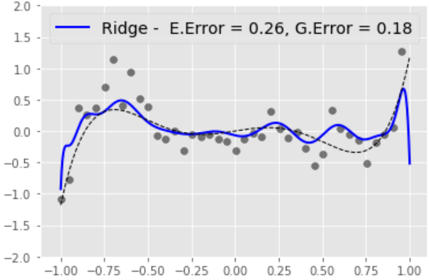
Polynomial Regression $(n > m)$

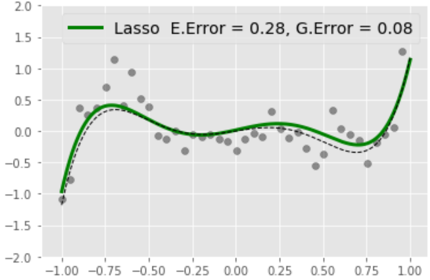
Linear Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Ridge Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Sparse Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$





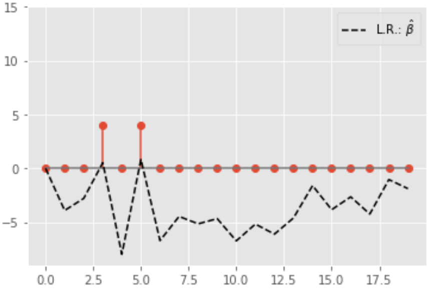


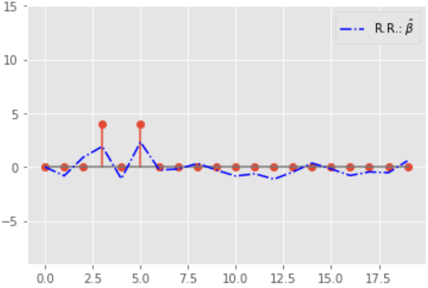
Polynomial Regression $(n > m)$

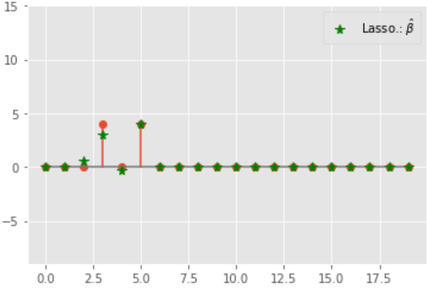
Linear Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Ridge Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Sparse Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$







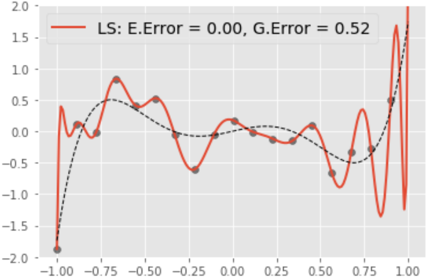
Polynomial Regression $(n < m)$ (!)

Polynomial Regression $(n < m)$ (!)

Linear Regression
$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

Ridge Regression
$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$

Sparse Regression
$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$

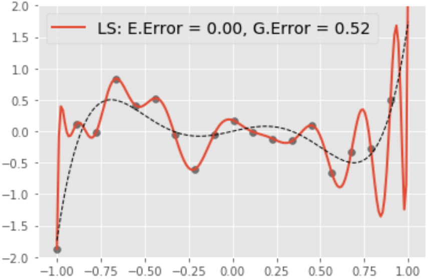


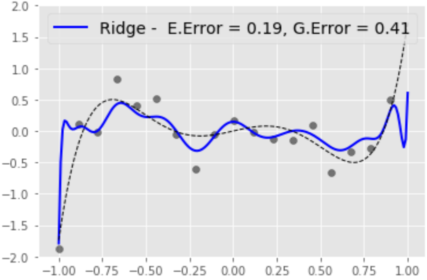
Polynomial Regression $(n < m)$ (!)

Linear Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Ridge Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Sparse Regression
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$





Polynomial Regression $(n < m)$ (!)

Linear Regression

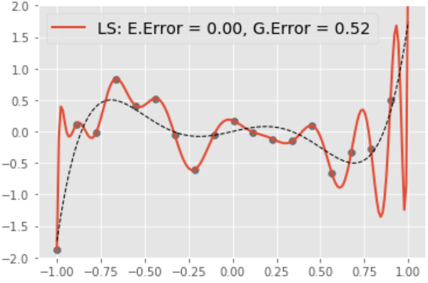
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

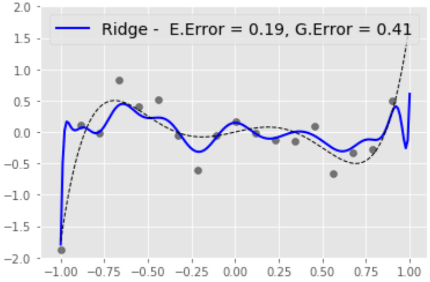
Ridge Regression

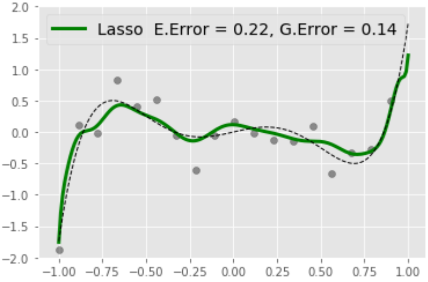
$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Sparse Regression

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$







Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$

Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$
- Use a linear model for the log-ratio of the probabilities:

$$\log \frac{P(Y=1|X=\mathbf{x})}{P(Y=0|X=\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$$

Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$
- Use a linear model for the log-ratio of the probabilities:

$$\log \frac{P(Y = 1|X = \mathbf{x})}{P(Y = 0|X = \mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$$

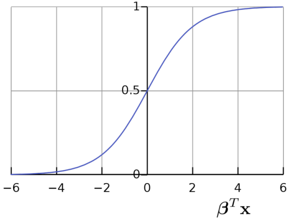
$$\Rightarrow P(Y = 1|X = \mathbf{x}) = \sigma(\boldsymbol{\beta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}}$$

Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$
- Use a linear model for the log-ratio of the probabilities:

$$\log \frac{P(Y = 1|X = \mathbf{x})}{P(Y = 0|X = \mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$$

$$\Rightarrow P(Y = 1|X = \mathbf{x}) = \sigma(\boldsymbol{\beta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}}$$

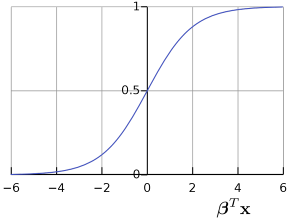


Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$
- Use a linear model for the log-ratio of the probabilities:

$$\log \frac{P(Y=1|X=\mathbf{x})}{P(Y=0|X=\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}$$

$$\Rightarrow P(Y=1|X=\mathbf{x}) = \sigma(\boldsymbol{\beta}^T \mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^T \mathbf{x}}}$$



Generalized Linear Models (GLM)

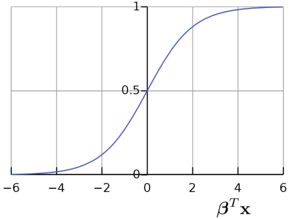
$$g(\mu(X)) = \boldsymbol{\beta}^T \mathbf{x}$$

Logistic Regression

- Classification (Binary): $\mathcal{Y} = \{0, 1\}$
- Goal: provide a hypothesis that approximates $p(Y|X)$
- Use a linear model for the log-ratio of the probabilities:

$$\log \frac{P(Y=1|X=\mathbf{x})}{P(Y=0|X=\mathbf{x})} = \boldsymbol{\beta}^T\mathbf{x}$$

$$\Rightarrow P(Y=1|X=\mathbf{x}) = \sigma(\boldsymbol{\beta}^T\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^T\mathbf{x}}}$$



Generalized Linear Models (GLM)

$$g(\mu(X)) = \boldsymbol{\beta}^T\mathbf{x}$$

For log. regression, $\mu(X) = P(Y=1|X=\mathbf{x})$, and $g(u) = \log \frac{u}{1-u}$

Logistic Regression

How do we fit $\boldsymbol{\beta}$? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

Logistic Regression

How do we fit $\boldsymbol{\beta}$? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} P(Y = 1|\mathbf{x}_i)^{y_i} P(Y = 0|\mathbf{x}_i)^{1-y_i}$$

Logistic Regression

How do we fit β? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} P(Y = 1|\mathbf{x}_i)^{y_i} P(Y = 0|\mathbf{x}_i)^{1-y_i}$$

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log P(Y = 1|\mathbf{x}_i) + (1 - y_i) \log P(Y = 0|\mathbf{x}_i)$$

Logistic Regression

How do we fit β? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} P(Y = 1|\mathbf{x}_i)^{y_i} P(Y = 0|\mathbf{x}_i)^{1-y_i}$$

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log P(Y = 1|\mathbf{x}_i) + (1 - y_i) \log P(Y = 0|\mathbf{x}_i)$$

$$= \sum_{i=1}^{n} y_i \log \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}} + (1 - y_i) \log \frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}}}$$

Logistic Regression

How do we fit β? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} P(Y = 1|\mathbf{x}_i)^{y_i} P(Y = 0|\mathbf{x}_i)^{1-y_i}$$

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log P(Y = 1|\mathbf{x}_i) + (1 - y_i) \log P(Y = 0|\mathbf{x}_i)$$

$$= \sum_{i=1}^{n} y_i \log \frac{1}{1 + e^{-\boldsymbol{\beta}^T\mathbf{x}}} + (1 - y_i) \log \frac{1}{1 + e^{\boldsymbol{\beta}^T\mathbf{x}}}$$

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} y_i \boldsymbol{\beta}^T\mathbf{x} - \log(1 + e^{\boldsymbol{\beta}^T\mathbf{x}_i})$$

Logistic Regression

How do we fit β? Maximize the (log) likelihood $P(\{\mathbf{y}_i\}|\boldsymbol{\beta}, \{\mathbf{x}_i\})$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y = y_i|\mathbf{x}_i) = \prod_{i=1}^{n} P(Y = 1|\mathbf{x}_i)^{y_i} P(Y = 0|\mathbf{x}_i)^{1-y_i}$$

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log P(Y = 1|\mathbf{x}_i) + (1 - y_i) \log P(Y = 0|\mathbf{x}_i)$$

$$= \sum_{i=1}^{n} y_i \log \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}} + (1 - y_i) \log \frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}}}$$

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} y_i \boldsymbol{\beta}^T \mathbf{x} - \log(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i})$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log(1 + e^{-\tilde{y}_i(\boldsymbol{\beta}^T \mathbf{x})}); \quad \tilde{y} = 2y - 1 \in \{\pm 1\}$$

Need for Regularization

Ridge ℓ_2 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

Need for Regularization

Ridge ℓ_2 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Need for Regularization

Ridge ℓ_2 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Leukemia Dataset (lymphoblastic vs myeloid) $n = 35, p = 7128$!

Need for Regularization

Ridge ℓ_2 regularization:

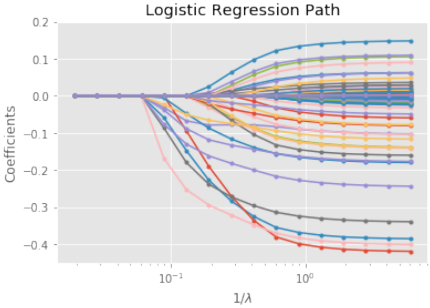
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Leukemia Dataset (lymphoblastic vs myeloid) $n = 35, p = 7128$!

- Sparse Logistic Regression



Need for Regularization

Ridge ℓ_2 regularization:

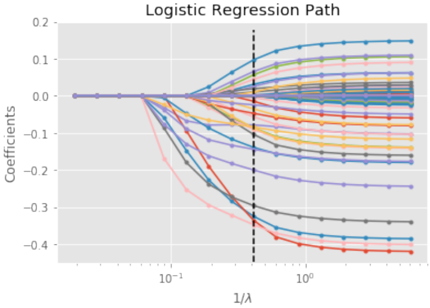
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Leukemia Dataset (lymphoblastic vs myeloid) $n = 35, p = 7128$!

- Sparse Logistic Regression



Need for Regularization

Ridge ℓ_2 regularization:

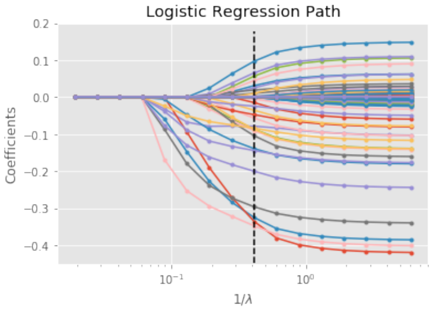
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

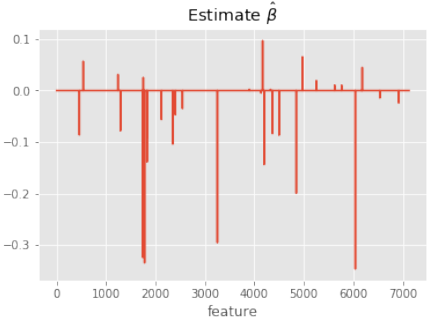
Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Leukemia Dataset (lymphoblastic vs myeloid) $n = 35, p = 7128$!

- Sparse Logistic Regression





Need for Regularization

Ridge ℓ_2 regularization:

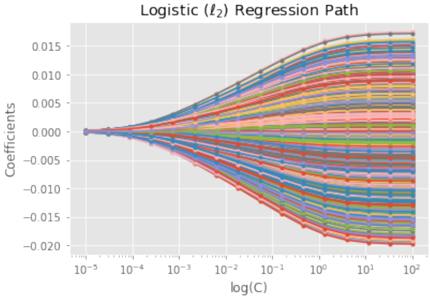
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

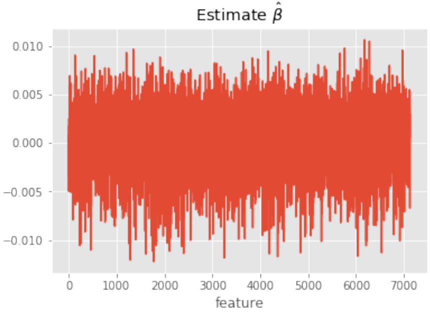
Sparse ℓ_1 regularization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -\log \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

Leukemia Dataset (lymphoblastic vs myeloid) $n = 35, p = 7128$!

- ℓ_2 Logistic Regression





Multi-class Logistic Regression

Multi-class Logistic Regression

- Multinomial likelihood

$$P(Y = k|\mathbf{x}) = \frac{e^{\boldsymbol{\beta}_k^T \mathbf{x}}}{\sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^T \mathbf{x}}}; \quad K \text{ classifiers } \boldsymbol{\beta}_l \in \mathbb{R}^m$$

Multi-class Logistic Regression

- Multinomial likelihood

$$P(Y = k|\mathbf{x}) = \frac{e^{\boldsymbol{\beta}_k^T \mathbf{x}}}{\sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^T \mathbf{x}}}; \quad K \text{ classifiers } \boldsymbol{\beta}_l \in \mathbb{R}^m$$

- Regularized log-likelihood

$$\frac{-1}{n} \sum_{i=1}^{n} \log P(Y = y_i|\mathbf{x}_i) + \lambda \sum_{l=1}^{K} \|\boldsymbol{\beta}_l\|_1$$

Multi-class Logistic Regression

- Multinomial likelihood

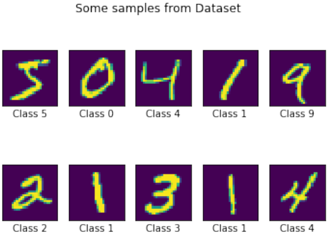
$$P(Y = k|\mathbf{x}) = \frac{e^{\boldsymbol{\beta}_k^T \mathbf{x}}}{\sum_{l=1}^{K} e^{\boldsymbol{\beta}_l^T \mathbf{x}}}; \quad K \text{ classifiers } \boldsymbol{\beta}_l \in \mathbb{R}^m$$

- Regularized log-likelihood

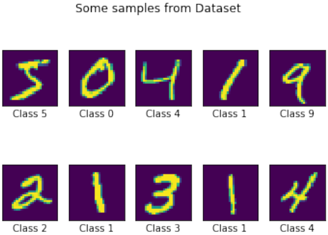
$$\frac{-1}{n} \sum_{i=1}^{n} \log P(Y = y_i|\mathbf{x}_i) + \lambda \sum_{l=1}^{K} \|\boldsymbol{\beta}_l\|_1$$

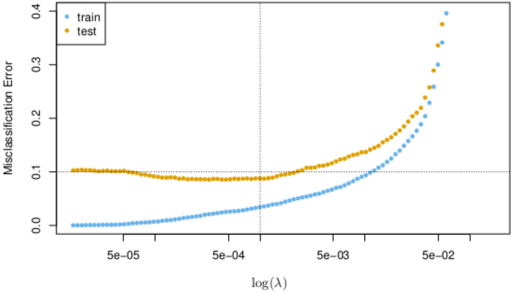
$$\frac{-1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} 1_{[y_i=k]} (\boldsymbol{\beta}_k^T \mathbf{x}_i) - \log \left(\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k^T \mathbf{x}_i} \right) \right] + \lambda \sum_{l=1}^{K} \|\boldsymbol{\beta}_l\|_1$$

Digits Classification

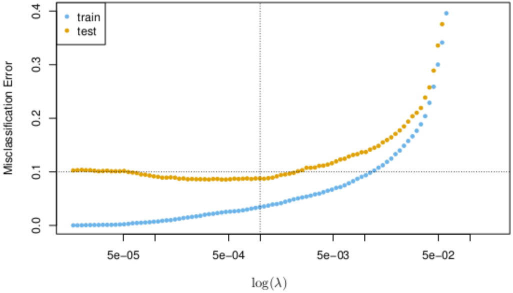


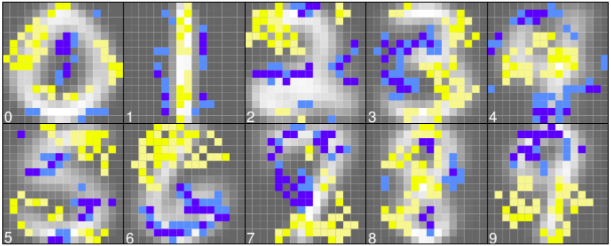
Digits Classification



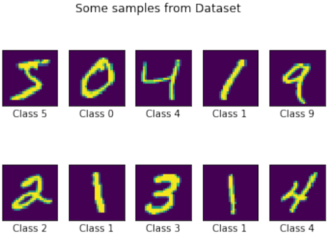


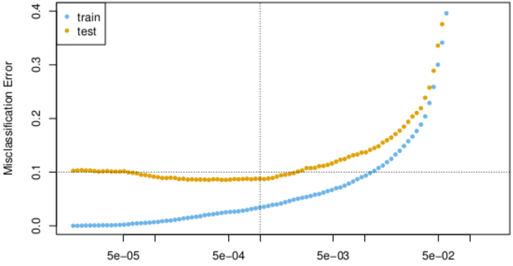
Digits Classification

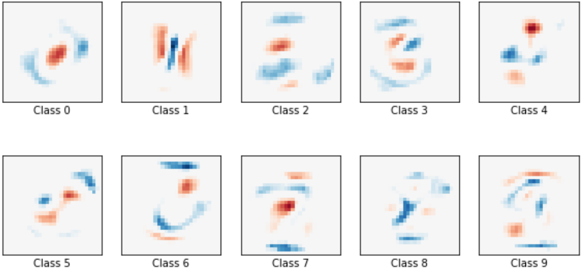




Digits Classification







Sparsity in Machine Learning
Part II

EN.580.709 - Fall 2019

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n \times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n\times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Consider $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{x}_m] \in \mathbb{R}^{n\times(m+1)}$, and

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n \times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Consider $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{x}_m] \in \mathbb{R}^{n \times (m+1)}$, and

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

As long as $\tilde{\boldsymbol{\beta}}_m + \tilde{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m$, loss is unchanged!

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n \times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Consider $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{x}_m] \in \mathbb{R}^{n \times (m+1)}$, and

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

As long as $\tilde{\boldsymbol{\beta}}_m + \tilde{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m$, loss is unchanged!

- Solution:

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n \times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Consider $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{x}_m] \in \mathbb{R}^{n \times (m+1)}$, and

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

As long as $\tilde{\boldsymbol{\beta}}_m + \tilde{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m$, loss is unchanged!

- Solution: penalize $\|\boldsymbol{\beta}\|_2^2 \Rightarrow \tilde{\boldsymbol{\beta}}_m = \tilde{\boldsymbol{\beta}}_{m+1} = \frac{\hat{\boldsymbol{\beta}}_m}{2}$ (why?)

Recall Lasso

- $\mathbf{y} \in \mathbb{R}^n$: response, $\mathbf{X} \in \mathbb{R}^{n \times m}$: predictors/features

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Consider $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{x}_m] \in \mathbb{R}^{n \times (m+1)}$, and

$$\tilde{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

As long as $\tilde{\boldsymbol{\beta}}_m + \tilde{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m$, loss is unchanged!

- Solution: penalize $\|\boldsymbol{\beta}\|_2^2 \Rightarrow \tilde{\boldsymbol{\beta}}_m = \tilde{\boldsymbol{\beta}}_{m+1} = \frac{\hat{\boldsymbol{\beta}}_m}{2}$ (why?)

Elastic-Net

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2\right)$$

Example

Let $Z_1, Z_2 \sim N(0, 1)$

Example

Let $Z_1, Z_2 \sim N(0,1)$

$Y = 3Z_1 - 1.5Z_2 + 2\epsilon$, with $\epsilon \sim N(0,1)$

Example

Let $Z_1, Z_2 \sim N(0,1)$

$Y = 3Z_1 - 1.5Z_2 + 2\epsilon$, with $\epsilon \sim N(0,1)$

$X_j = Z_1 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 1, 2, 3$, and

Example

Let $Z_1, Z_2 \sim N(0,1)$

$Y = 3Z_1 - 1.5Z_2 + 2\epsilon$, with $\epsilon \sim N(0,1)$

$X_j = Z_1 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 1,2,3$, and

$X_j = Z_2 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 4,5,6$,

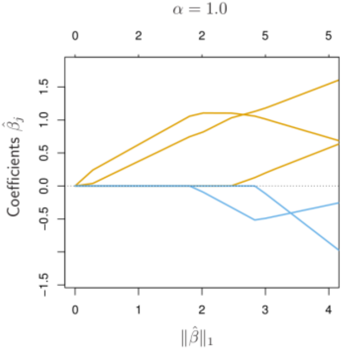
Example

Let $Z_1, Z_2 \sim N(0,1)$

$Y = 3Z_1 - 1.5Z_2 + 2\epsilon$, with $\epsilon \sim N(0,1)$

$X_j = Z_1 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 1,2,3$, and

$X_j = Z_2 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 4,5,6$,



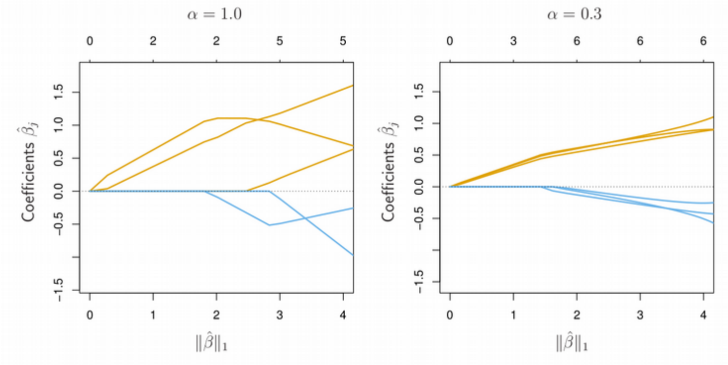
Example

Let $Z_1, Z_2 \sim N(0,1)$

$Y = 3Z_1 - 1.5Z_2 + 2\epsilon$, with $\epsilon \sim N(0,1)$

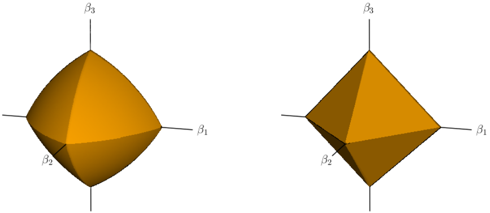
$X_j = Z_1 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 1,2,3$, and

$X_j = Z_2 + \zeta_j/5$, with $\zeta_j \sim N(0,1)$, for $j = 4,5,6$,



The Elastic-Net Problem is Strictly Convex

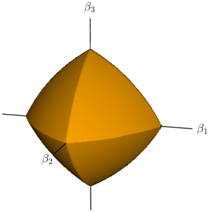
$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$$



(why?)

The Elastic-Net Problem is Strictly Convex

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$$



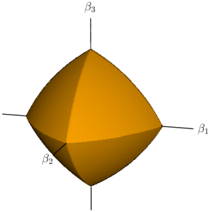
(why?)

How to optimize?

- If $\mathbf{X}$: orthogonal $\Rightarrow$

The Elastic-Net Problem is Strictly Convex

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$$



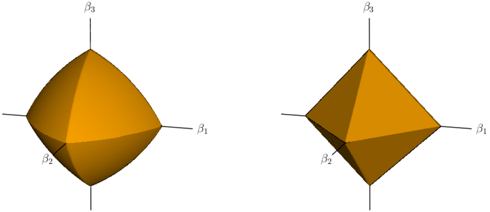
(why?)

How to optimize?

- If $\mathbf{X}$: orthogonal $\Rightarrow \hat{\boldsymbol{\beta}} = \frac{1}{1+\alpha} S_\lambda(\mathbf{X}^T\mathbf{y})$
- If $\mathbf{X}$: redundant?

The Elastic-Net Problem is Strictly Convex

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$$



(why?)

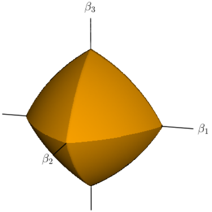
How to optimize?

- If $\mathbf{X}$: orthogonal $\Rightarrow \hat{\boldsymbol{\beta}} = \frac{1}{1+\alpha} S_\lambda(\mathbf{X}^T\mathbf{y})$
- If $\mathbf{X}$: redundant?

$$\boldsymbol{\beta}^{k+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(\boldsymbol{\beta}^k - \eta\nabla f(\boldsymbol{\beta}^k)\right)$$

The Elastic-Net Problem is Strictly Convex

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$$



(why?)

How to optimize?

- If $\mathbf{X}$: orthogonal $\Rightarrow \hat{\boldsymbol{\beta}} = \frac{1}{1+\alpha} S_\lambda(\mathbf{X}^T\mathbf{y})$
- If $\mathbf{X}$: redundant?

$$\boldsymbol{\beta}^{k+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(\boldsymbol{\beta}^k - \eta\nabla f(\boldsymbol{\beta}^k)\right) = S_{\lambda/c}\left(\boldsymbol{\beta}^k - \frac{1}{c}[\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^k - \mathbf{y}) + \alpha\boldsymbol{\beta}^k]\right)$$

Leukemia type (AML/ALL) dataset - Golub et al, Science '90

- 38 training samples, 34 test samples, $m = 7129$ genes.
- Record the expression for sample i and gene j

Leukemia type (AML/ALL) dataset - Golub et al, Science '90

- 38 training samples, 34 test samples, $m = 7129$ genes.
- Record the expression for sample i and gene j

Leukemia classification example

Method	10-fold CV error	Test error	No. of genes
Golub UR	3/38	4/34	50
SVM RFE	2/38	1/34	31
PLR RFE	2/38	1/34	26
NSC	2/38	2/34	21
Elastic Net	2/38	0/34	45

UR: univariate ranking (Golub et al. 1999)
RFE: recursive feature elimination (Guyon et al. 2002)
SVM: support vector machine (Guyon et al. 2002)
PLR: penalized logistic regression (Zhu and Hastie 2004)
NSC: nearest shrunken centroids (Tibshirani et al. 2002)

[Hui & Hasie, '05]

Group Lasso

- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \ldots, \beta_{j,m_j}]$

Group Lasso

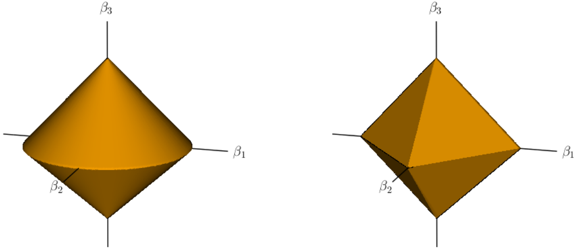
- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \ldots, \beta_{j,m_j}]$

$$\min_{\boldsymbol{\beta}_j} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$$

Group Lasso

- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \dots, \beta_{j,m_j}]$

$$\min_{\boldsymbol{\beta}_j} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$$



Group Lasso

- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \ldots, \beta_{j,m_j}]$

$$\min_{\boldsymbol{\beta}_j} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$$

- What if we have *different tasks* to predict [i.e. multi-variate regression]?
- Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_J]$: responses

Group Lasso

- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \ldots, \beta_{j,m_j}]$

$$\min_{\boldsymbol{\beta}_j} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$$

- What if we have *different tasks* to predict [i.e. multi-variate regression]?
- Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_J]$: responses

$$\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \lambda \|\mathbf{B}\|_{1,2}; \quad \|\mathbf{B}\|_{1,2} = \sum_{i=1}^{m} \|\mathbf{B}_{i,:}\|_2$$

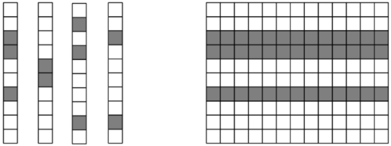
Group Lasso

- What if there's some natural grouping in the problem, and we want to introduce this as a prior?
- Say, J groups, and each $\boldsymbol{\beta}_j = [\beta_{j,1}, \ldots, \beta_{j,m_j}]$

$$\min_{\boldsymbol{\beta}_j} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$$

- What if we have *different tasks* to predict [i.e. multi-variate regression]?
- Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_J]$: responses

$$\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_2^2 + \lambda \|\mathbf{B}\|_{1,2}; \quad \|\mathbf{B}\|_{1,2} = \sum_{i=1}^{m} \|\mathbf{B}_{i,:}\|_2$$



Sparsity vs. joint sparsity

End of this Part

Where are we at?

Where are we at?

1. Sparsity in linear systems of equations

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \;\; \text{s.t.} \;\; \|\mathbf{x}\|_0 \leq k$$

2. Basis Pursuit and Compressed Sensing

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

Where are we at?

1. Sparsity in linear systems of equations

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \text{ s.t. } \|\mathbf{x}\|_0 \leq k$$

2. Basis Pursuit and Compressed Sensing

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

3. Sparsity in (linear, logistic, multi-variate) regression

$$\min_{\boldsymbol{\beta}} \sum_{i}^{n} \mathcal{L}(y_i, \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_p$$

Where are we at?

1. Sparsity in linear systems of equations

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \ \text{ s.t. } \ \|\mathbf{x}\|_0 \leq k$$

2. Basis Pursuit and Compressed Sensing

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

3. Sparsity in (linear, logistic, multi-variate) regression

$$\min_{\boldsymbol{\beta}} \sum_{i}^{n} \mathcal{L}(y_i, \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_p$$

5. Dictionary Learning:

$$\min_{\mathbf{x},\mathbf{A}} \sum_{i}^{n} \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \ \text{ s.t. } \ \|\mathbf{x}_i\|_0 \leq k, \forall i$$

Where are we at?

1. Sparsity in linear systems of equations

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \ \text{ s.t. } \ \|\mathbf{x}\|_0 \leq k$$

2. Basis Pursuit and Compressed Sensing

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

3. Sparsity in (linear, logistic, multi-variate) regression

$$\min_{\boldsymbol{\beta}} \sum_i^n \mathcal{L}(y_i, \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_p$$

4. Matrix spectral sparsity and robust PCA:

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{A}\| + \lambda \phi(\mathbf{A}) \qquad \longleftarrow$$

5. Dictionary Learning:

$$\min_{\mathbf{x}, \mathbf{A}} \sum_i^n \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2^2 \ \text{ s.t. } \ \|\mathbf{x}_i\|_0 \leq k, \forall i$$

Matrix Spectral Sparsity

EN.580.709 - Fall 2019

Eigenfaces

- Yale B Dataset

Eigenfaces

- Yale B Dataset

- Factoring Illumination

(a) mean face (b) first eigenface (c) second eigenface

Eigenfaces

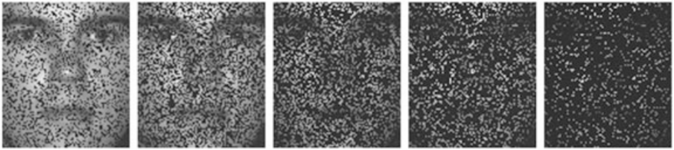
- Yale B Dataset

- Factoring Illumination

(a) mean face (b) first eigenface (c) second eigenface

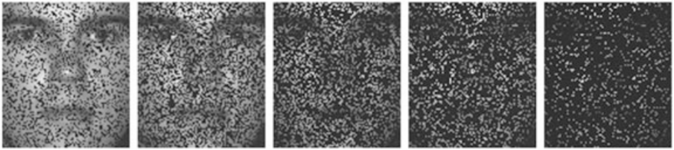
(a) Variation along the first eigenface

Matrix Completion

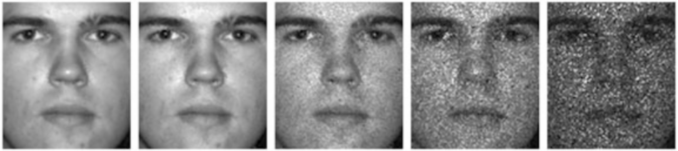


(a) Face images with (30, 50, 70, 80, 90)% percentage of missing entries

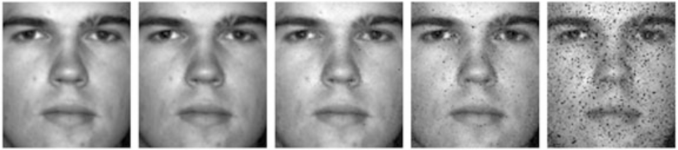
Matrix Completion



(a) Face images with (30, 50, 70, 80, 90)% percentage of missing entries

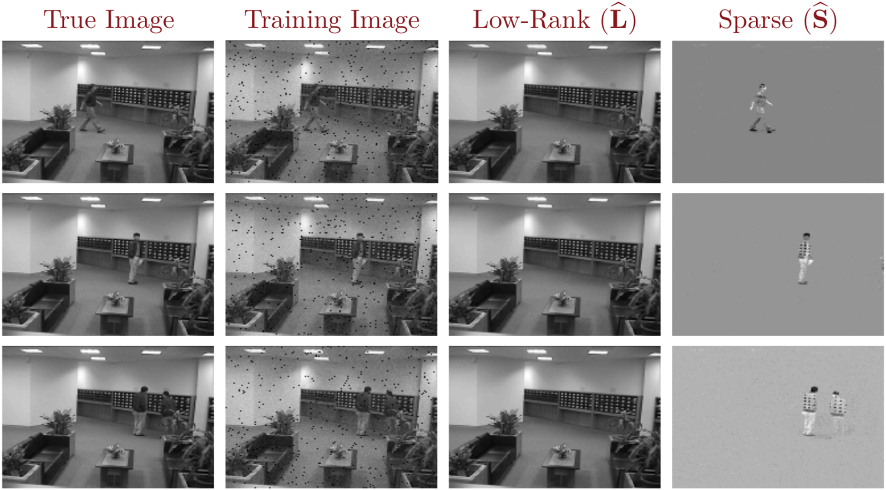


(c) Face images reconstructed by convex optimization with $\tau = 2 \times 10^4$



(d) Face images reconstructed by convex optimization with $\tau = 4 \times 10^5$

Robust PCA


True Image
Training Image
Low-Rank $(\widehat{\mathbf{L}})$
Sparse $(\widehat{\mathbf{S}})$