CSE 6740 A/ISyE 6740: Computational Data Analysis: Introductory lecture

Nisha Chandramoorthy

August 31, 2023

Last time

- Least squares regression

Last time

- Least squares regression
- Gauss-Markov theorem

Last time

- Least squares regression
- Gauss-Markov theorem
- Ridge regression, optimization and geometric perspectives

Last time

- Least squares regression
- Gauss-Markov theorem
- Ridge regression, optimization and geometric perspectives
- Shrinkage

Today

- Shrinkage by ridge regression

Today

- Shrinkage by ridge regression
- Geometric view of LASSO

Today

- Shrinkage by ridge regression
- Geometric view of LASSO
- Generalization of LASSO

Today

- Shrinkage by ridge regression
- Geometric view of LASSO
- Generalization of LASSO
- ℓ^0 regularization and compressed sensing

Today

- Shrinkage by ridge regression
- Geometric view of LASSO
- Generalization of LASSO
- ℓ^0 regularization and compressed sensing

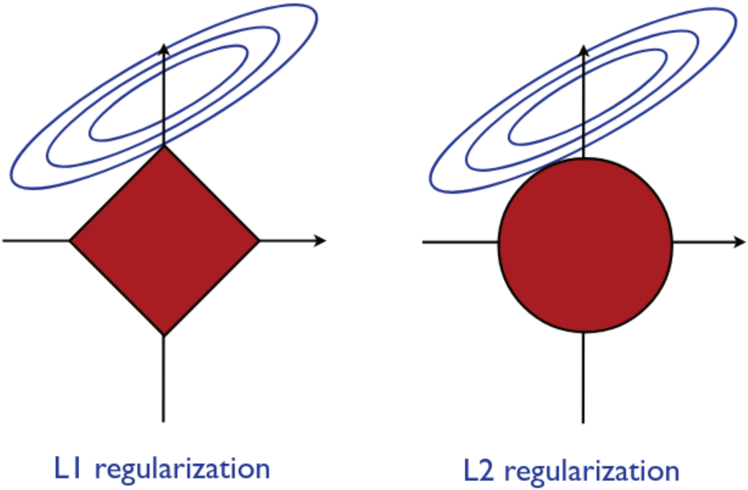
- Online linear regression algorithms

- Online linear regression algorithms
- Classification ERM

- Online linear regression algorithms
- Classification ERM
- Perceptron algorithm

- Online linear regression algorithms
- Classification ERM
- Perceptron algorithm
- Convergence proof of perceptron algorithm

Compression by LASSO



Shrinkage (Ridge) Algorithms

$$\mathcal{H} := \{ w^T x + b \; : \; w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

$z = (x, y)$

$$\ell(z, h) = (w^T x + b - y)^2$$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} (w^T x_i + b - y_i)^2$$

(w, b) $\quad (x_i, y_i) \in S$

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \hat{R}_S^{OLS}(h) + \lambda \|w\|^2$$

$\parallel$

(w^*, b^*)

$$\min_w \|Xw - Y\|^2 + \lambda \|w\|^2$$

$$w^{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

$$\underset{m \times d}{X} = \begin{bmatrix} \overline{\quad x_1^T \quad} \\ \vdots \\ \overline{\quad x_m^T \quad} \end{bmatrix} \qquad \underset{m}{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$X = U \Sigma V^T \qquad U = [u_1 | u_2 | \cdots | u_d]$$

$$V = [v_1 | v_2 | \cdots | v_d]$$

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{bmatrix}$$

$Y^{ridge}_{pred} =$

$$X w^{ridge} = X (X^T X + \lambda I)^{-1} X^T Y$$

$$= U \Sigma V^T (V \Sigma V^T + \lambda I)^{-1} V \Sigma U^T Y$$

$V^T = V^{-1}$

$$A^{-1} B^{-1} = (BA)^{-1}$$

$$= \sum_{i=1}^{d} u_i \underbrace{\frac{\sigma_i^2}{\lambda + \sigma_i^2}}_{} u_i^T Y$$

Shrinkage $\uparrow$ when $\sigma_i \downarrow$.

Convex quadratic program

$\rightarrow$ QP

$\rightarrow$ iterative methods

$$\hat{R}_S(w) = \frac{1}{m} \|\underline{X} w - \underline{Y}\|^2 + \lambda \|w\|^2$$

<u>Gradient</u> descent

$$w^{(t+1)} = w^{(t)} - \eta \nabla \hat{R}_S(w^{(t)})$$

$\uparrow$ learning rate

Stochastic <u>Gradient</u> descent (SGD)

$$w^{(t+1)} = w^{(t)} - \eta \nabla \tilde{\hat{R}}_S(w^{(t)})$$

Batch $b < m$ samples

Best subset selection

$$\hat{R}_S(\omega) = \|X\omega - Y\|^2 + \lambda \|\omega\|_0 \leftarrow$$

$\|\omega\|_0$: ℓ^0 norm   Tibshirani 2000s

$\|$

\# of nonzero entries of ω.

$$\begin{bmatrix} x_1^T \cdots \\ \vdots \cdots \\ x_m^T \cdots \end{bmatrix} \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_d \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

short wide

X

$m \ll d$

$$\|Y - XW\|$$

Optimization way of seeing Compression in the LASSO

in the case of X orthonormal

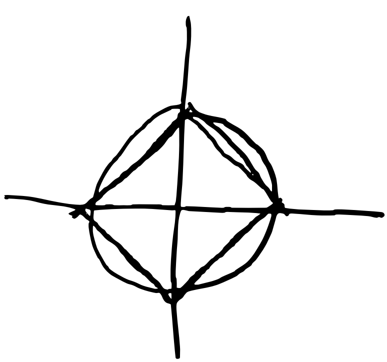
$$\underbrace{\hat{R}_S(\omega)} = \| X\omega - Y \|^2 + \lambda \|\omega\|_1$$

$$\|\omega\|_1 = \sum_{j=1}^{d} |\omega_j|$$

LASSO estimator

ERM objective

$\hat{R}_S^{lasso}$, $\hat{R}_S^{bss}$, $\hat{R}_S^{ridge}$

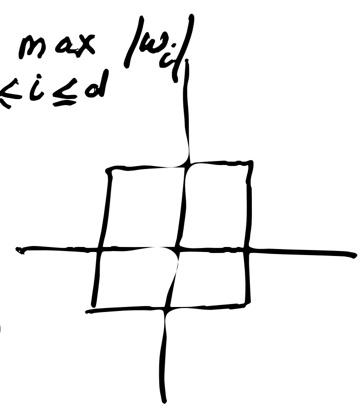


$\|\omega\|_\infty = \gamma$ $\quad \max_{1 \le i \le d} |\omega_i|$

$\|\omega\| = \gamma$

$\|\omega\|_1 = \gamma$

$|\omega_1| + |\omega_2|$



$\rightarrow \quad X w \approx Y$

$m \ll d$

Candes, Romberg, Tao 2005

$\rightarrow \quad X\underline{w} = Y + \varepsilon$

Signal processing, image processing

$$S_j = \sum_{i=1}^{d} x_{ji} w_i$$

Stable recovery of sparse w.

[Candes, Romberg, Tao 2005]

[Candes, Tao 2004]

Thm: $\arg\min_{w} \|w\|_1$ s.t. $\boxed{Xw = Y}$ ←

is the exact solution of $X\underset{\sim}{w} = Y$

for any true sparse w if

$nnz(w) < S$, where X satisfies S-restricted isometry. $\|w\|_0$

$T \subseteq \{1, 2, 3, \dots, d\}$

That is,

for all subsets of indices with $|T| < S$, there is $\delta_S > 0$ s.t.

(i) $(1-\delta_S)\|v\|^2 < \|X_T v\|^2 \leq (1+\delta_S)\|v\|^2$

(ii) $\delta_S + \delta_{2S} + \delta_{3S} < 1$ ←

$X = [\dots | \dots]$

X: Gaussian

Fourier basis

Stable recovery

Candes Romberg Tao 2005

Let $w^{lasso} :=$

$$\arg\min \|w\|_1 \quad s.t. \quad \|Xw - Y\| \leq \boxed{\varepsilon}$$

Let true $\underline{w}$ be sparse with $nnz(w) \leq S$ $\|w\|_0$

with S s.t. $\underbrace{\delta_{3S} + 3\delta_{4S}}_{S\text{-isometry}} < 2$.

Then,

$$\|w^{lasso} - w\| \leq C_S \underline{\underline{\varepsilon}}$$

$$w^{ols} = (X^T X)^{-1} X^T Y$$

$$w^{ols} - w$$
$$= (X^T X)^{-1} X^T Y - (X^T X)^{-1} X^T X w$$
$$= (X^T X)^{-1} X^T (Y - Xw)$$

$$\|w^{ols} - w\| \leq C \|X^T \varepsilon\| \approx C\varepsilon$$

Compressed sensing

$\underline{m \ll d}$

→ $(X^T X)^{+}$ is computationally expensive $O(d^3)$

→ Interpretability

Generalization of regression

→ Hoeffding's inequality: $S_n = X_1 + X_2 + \dots + X_n$

$X_i \perp\!\!\!\perp X_j \quad 0 \le X_i \le L$

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq e^{-\frac{2t^2}{nL^2}}$$

Thm: $\sup_{z,h} \ell(z,h) = L$. Let $\mathcal{H}$ be finite.

Then, for every $\delta > 0$, with probability at least $\underline{1-\delta}$, $\forall\, h \in \mathcal{H}$,

$$R(h) \leq \hat{R}_S(h) + L\sqrt{\frac{\log|\mathcal{H}| + \log 1/\delta}{2m}}$$

Proof: Use Hoeffding's inequality

Generalization for Ridge & Lasso

$\mathcal{H}$ Complexity

Ridge

$$R(h) \le \hat{R}_S(h) + 4L\sqrt{\frac{\boxed{r^2\Lambda^2}}{m}} + L^2\sqrt{\frac{\log 1/\delta}{2m}}$$

where $r^2 \geq \Phi^T(x)\Phi(x) \;\; \forall x$

$\|w\| < \Lambda$

$x^T x \quad \|x\|^2$

Lasso $\mathcal{H} := \{ h_{w,b}(x) : w \in \mathbb{R}^d, b \in \mathbb{R}, \|w\|_1 < \Lambda \}$

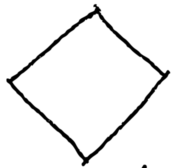
$$R(h) \le \hat{R}_S(h) + 2r_\infty \Lambda_1 L \sqrt{\frac{\boxed{2\log(2d)}}{m}} + L^2\sqrt{\frac{\log 1/\delta}{2m}}$$

where $\|X_i\|_\infty \le r_\infty$,

$\|w\|_1 \le \Lambda_1$

and $|h(x) - y| \le L$

$$\| XW - Y \|^2$$



$$X^T X = Id$$

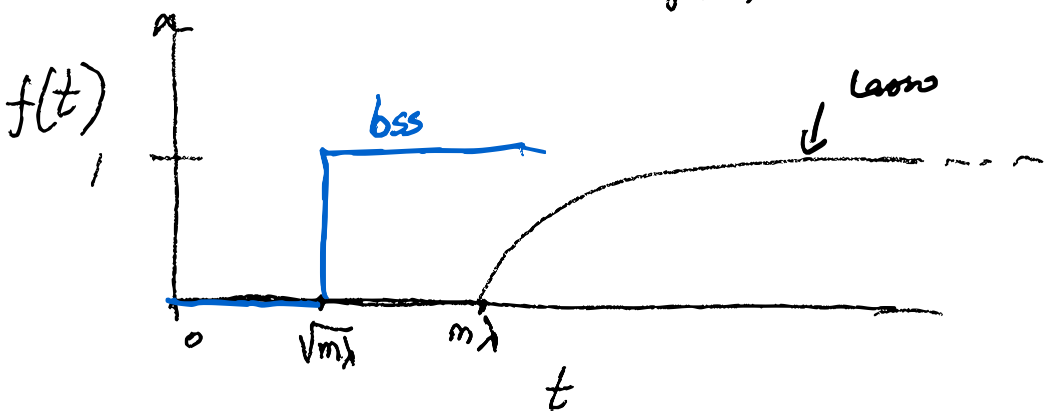
Subgradient

Proximal-descent

1) $$w^{OLS} = X^T Y$$

2) $$w^{ridge} = (X^T X + \lambda)^{-1} X^T Y$$
$$= (I + \lambda)^{-1} X^T Y$$

3) $$w_j^{lasso} = w_j^{ols} \max\left(0, 1 - \frac{m\lambda}{|w_j^{OLS}|}\right)$$
$$= f(t)$$



4) $$w_j^{bss} = w_j^{OLS} \, \mathbb{1}_{\{|w_j^{OLS}| > \sqrt{m\lambda}\}}$$