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

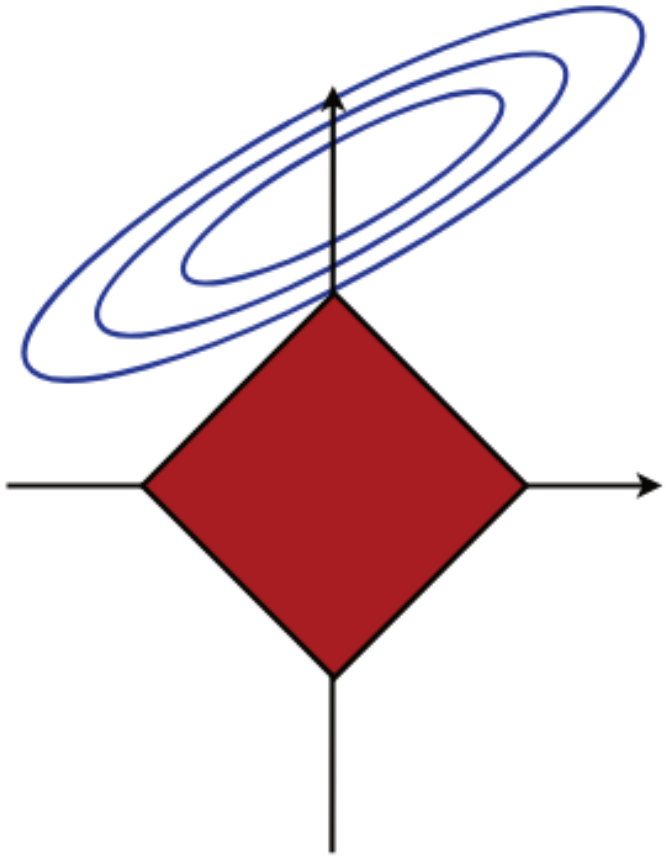
- Online linear regression algorithms

- Online linear regression algorithms
- Classification ERM

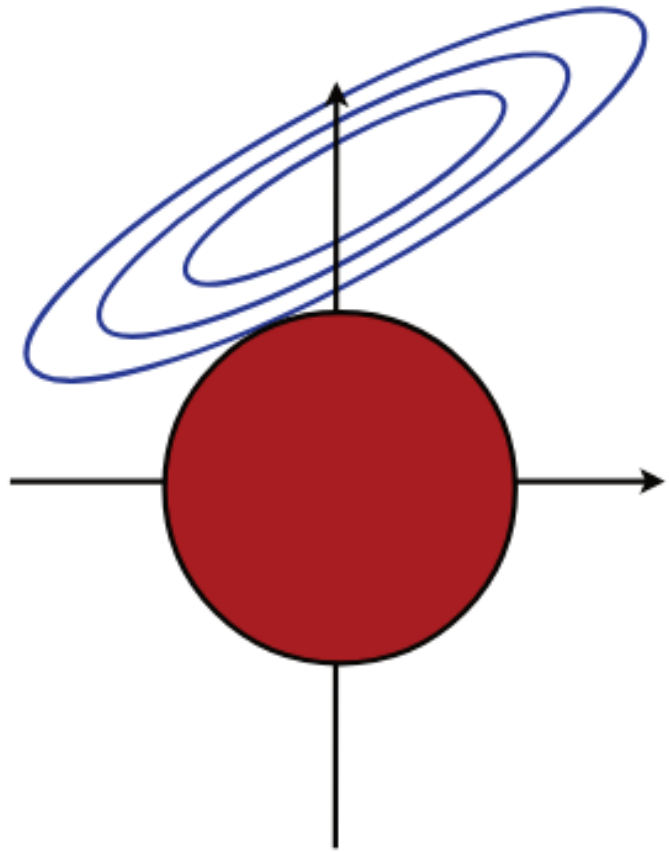
- Online linear regression algorithms
- Classification ERM
- Perceptron algorithm

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

Compression by LASSO



L1 regularization



L2 regularization

Shrinkage (Ridge) Algorithms

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

$z = (x, y)$

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

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

$h \downarrow (\omega, b)$

$(x_i, y_i) \in S$

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

$\omega^* \parallel (\omega^*, b^*)$

$$\min_{\omega} \ \|X\omega - Y\|^2 + \lambda \|\omega\|^2$$

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

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

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

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

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

$$Y_{pred}^{ridge} =$$

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

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

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

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

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

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

Convex quadratic program

→ QP

→ iterative methods

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

Gradient descent

$$\omega^{(t+1)} = \omega^{(t)} - \underbrace{\eta}_{\text{learning rate}} \nabla \hat{R}_S(\omega^{(t)})$$

Stochastic Gradient descent (SGD)

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

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

$\|\omega\|_0$ = # of nonzero entries of ω.

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

short wide

X

$m \ll d$

$$\|Y - XW\|$$

Optimization way of seeing Compression in the LASSO

in the case of X orthonormal

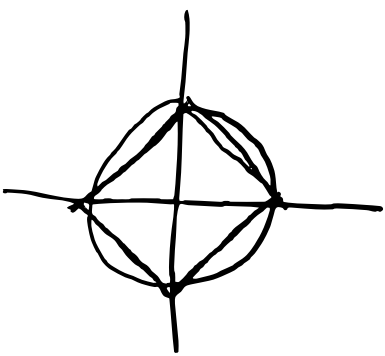
$$\hat{R}_S(w) = \|Xw - Y\|^2 + \lambda \|w\|_1$$

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

LASSO estimator

ERM objective

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

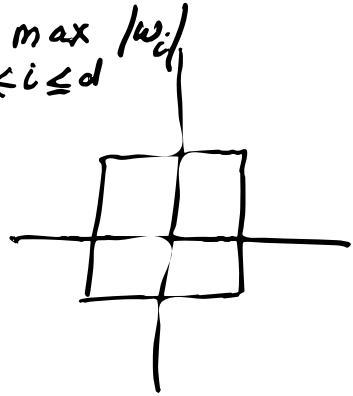


$\|w\|_\infty = \gamma \quad \max_{1 \le i \le d} |w_i|$

$\|w\| = \gamma$

$\|w\|_1 = \gamma$

$|w_1| + |w_2|$



→ $X \omega \approx Y$

$m \ll d$

→ Candes, Romberg, Tao 2005

$X\underline{\omega} = Y + \varepsilon$

Signal processing, image processing

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

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 $Xw = Y$

for any true sparse w if

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

$\|w\|_0$

That is,

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

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

$X = [\ldots | \ldots]$

X: Gaussian

Fourier basis

Stable recovery

Candes Romberg Tao 2005

Let $w^{lasso} :=$

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

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

with S s.t. $\delta_{3S} + 3\delta_{4S} < 2$.

(δ_{3S}, δ_{4S}: S-isometry)

Then,

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

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

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

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

Compressed sensing

$$m \ll d$$

→ $(X^T X)^{\dagger}$ is computationally expensive

$O(d^3)$

→ Interpretability

Generalization of regression

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

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

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

Thm: $\sup_{z,h} \ell(z,h) = L$. Let $\mathcal{H}$ be finite.
Then, for every $\delta > 0$, with probability at least $1-\delta$, $\forall\, h \in \mathcal{H}$,

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

Proof: Use Hoeffding's inequality

Generalization for Ridge & Lasso

$\mathcal{H}$ Complexity

Ridge

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

where $r^2 \geq \Phi^T(x)\Phi(x) \;\forall\; x$ $\quad \Phi \to x^Tx \quad \|x\|^2$

$\|w\| < \Lambda$

Lasso $\mathcal{H} := \{ h_{w,b}(x) : w \in \mathbb{R}^d, b \in \mathbb{R}, \|w\|_1 < 1 \}$

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

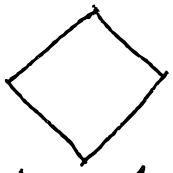
where $\|X_i\|_\infty \leq r_\infty$,

$\|w\|_1 \leq \Lambda_1$

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

$$\|Xw - Y\|^2$$

$$X^T X = Id$$



Subgradient
Proximal-descent

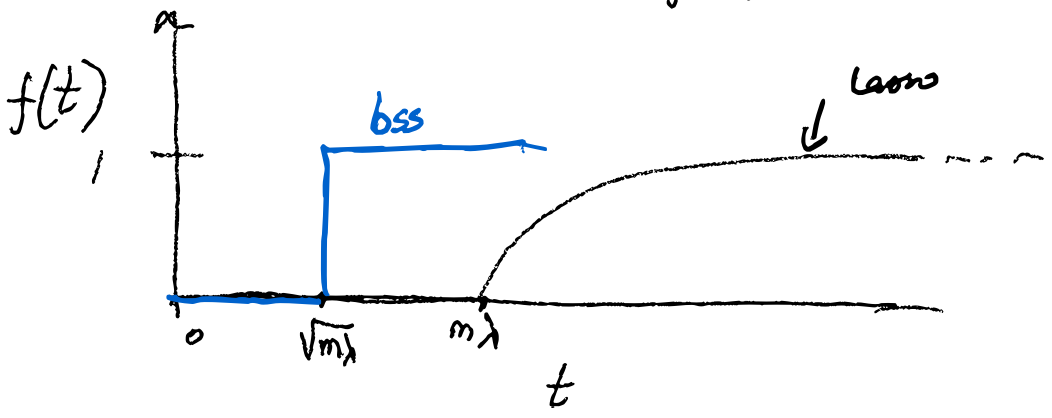
1) $w^{OLS} = X^T Y$

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

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

3) $w_j^{lasso} = w_j^{OLS} \max\left(0, 1 - \frac{n\lambda}{w_j^{OLS}}\right)$

$f(t)$



4) $w_j^{bss} = w_j^{OLS} \mathbb{1}\{w_j^{OLS} > \sqrt{n\lambda}\}$