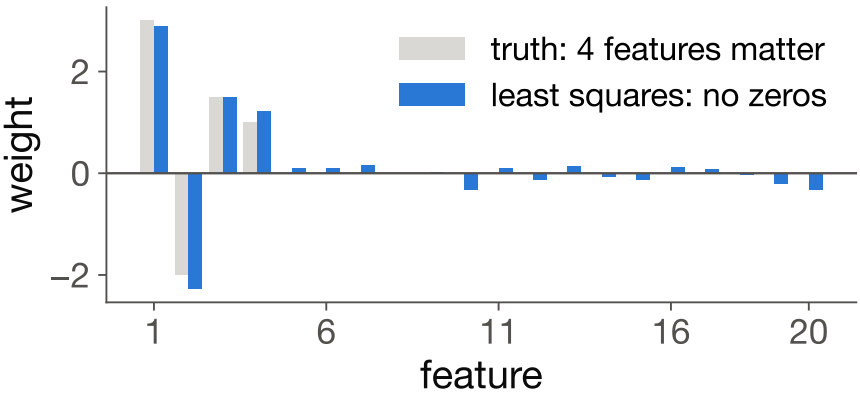
Proximal gradient and the lasso

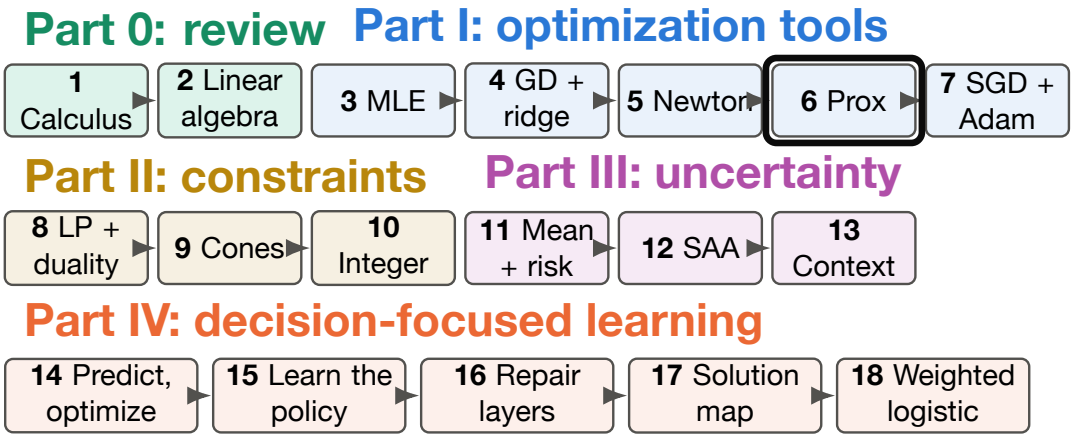
208424 Lecture 6

Only 4 of 20 features matter

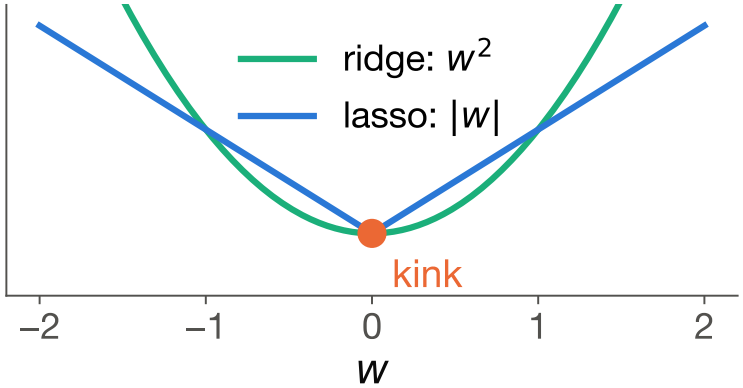


least squares gives all 20 features a nonzero weight

Course map


Part 0: review
Part I: optimization tools
1 Calculus
2 Linear algebra
3 MLE
4 GD + ridge
5 Newton
6 Prox
7 SGD + Adam
Part II: constraints
Part III: uncertainty
8 LP + duality
9 Cones
10 Integer
11 Mean + risk
12 SAA
13 Context
Part IV: decision-focused learning
14 Predict, optimize
15 Learn the policy
16 Repair layers
17 Solution map
18 Weighted logistic

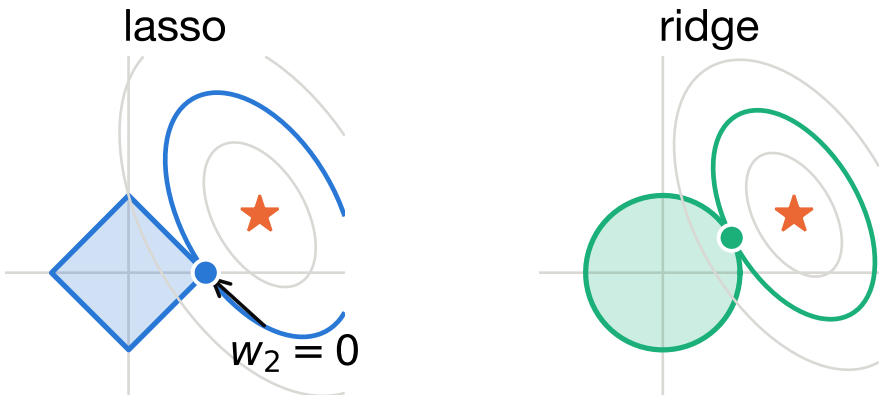
The lasso: a penalty with a kink



$$\min_w \ \tfrac{1}{2}\|y - Xw\|^2 + \lambda\|w\|_1, \quad \|w\|_1 = |w_1| + \cdots + |w_p|$$

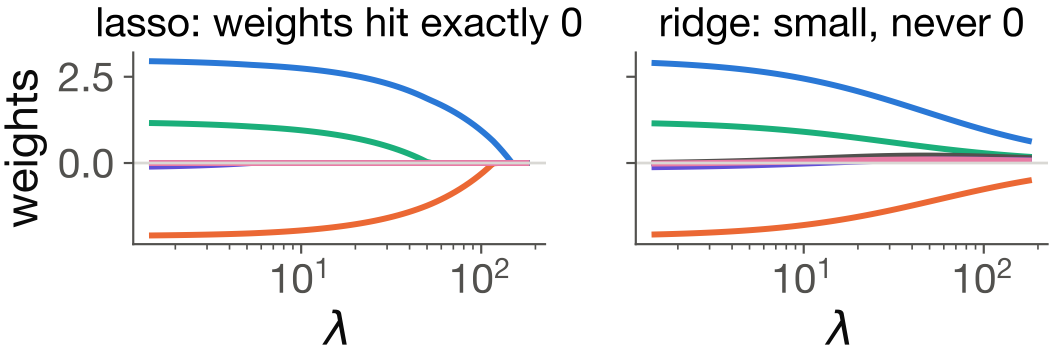
$\lambda \geq 0$: how much we penalize the weights

Why the lasso gives exact zeros



the loss touches the ℓ_1 ball at a vertex: $w_2 = 0$

Lasso path versus ridge path



here, larger λ drops the lasso's features one by one

Summary: the lasso

- Let X hold the p features as columns, y the outcomes, w the weights, $\lambda \geq 0$.
- Let $\|w\|_1 = |w_1| + \cdots + |w_p|$ and $\|w\|^2 = w_1^2 + \cdots + w_p^2$.
- We want: $\min_w \frac{1}{2}\|y - Xw\|^2 + \lambda\|w\|_1$, the lasso.
- The kink (sharp corner) of $|w|$ at 0: larger λ usually sets more w_j exactly to 0.
- Ridge uses $\lambda\|w\|^2$: no kink, small weights, never 0.

Checkpoint 1

$$w = (3,\ -2,\ 0)$$

$$\|w\|_1 = ? \qquad \|w\|^2 = ?$$

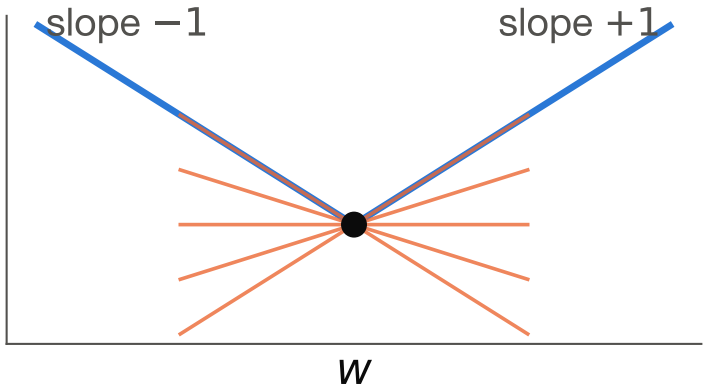
Checkpoint 1

$$w = (3,\ -2,\ 0)$$

$$\|w\|_1 = ? \qquad \|w\|^2 = ?$$

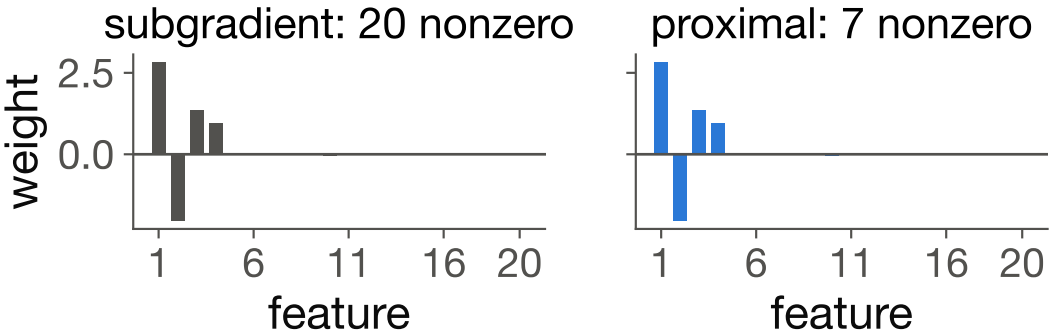
$$3 + 2 + 0 = 5 \qquad 9 + 4 + 0 = 13$$

No derivative at the kink: subgradients



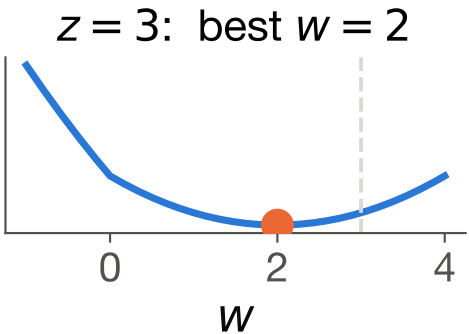
at 0: every slope in $[-1, 1]$ is a subgradient

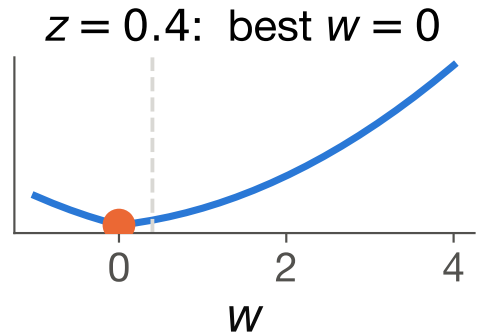
Without a proximal step: no exact zeros



3000 steps: the proximal step gives exact zeros

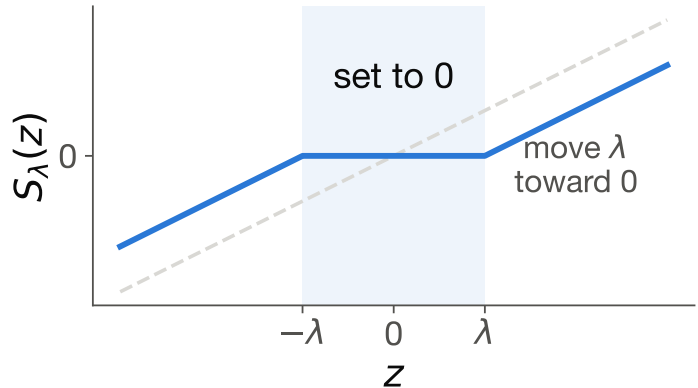
A tiny problem we can solve exactly





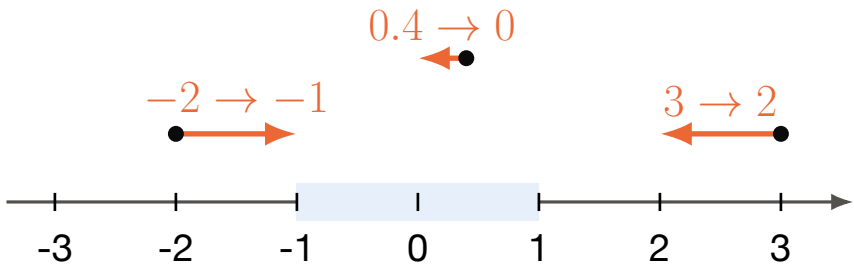
$$\min_w \ \tfrac{1}{2}(w-z)^2 + \lambda|w| \qquad (\lambda = 1)$$

The answer: soft-thresholding



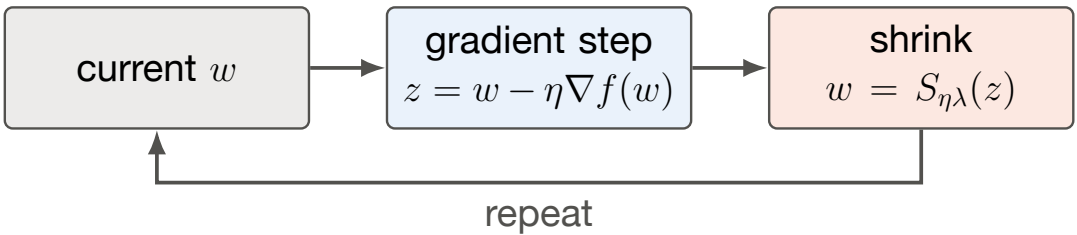
$$S_\lambda(z) = \text{sign}(z)\ \max\left(|z| - \lambda,\ 0\right)$$

Soft-thresholding with numbers



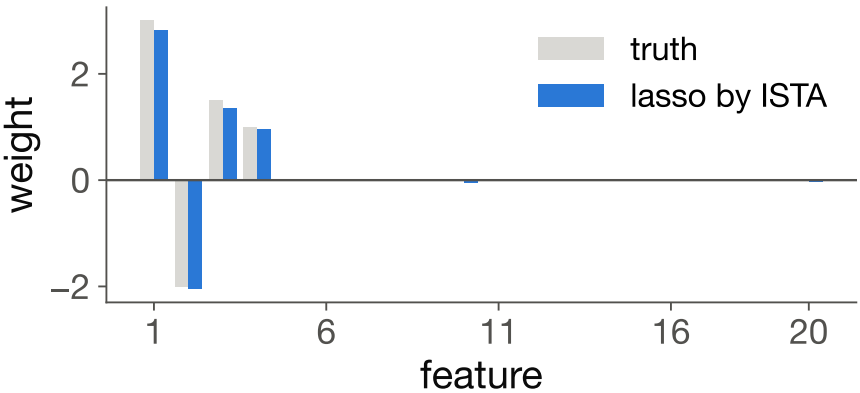
$\lambda = 1$: move every number 1 toward 0, stop at 0

ISTA: gradient step, then shrink



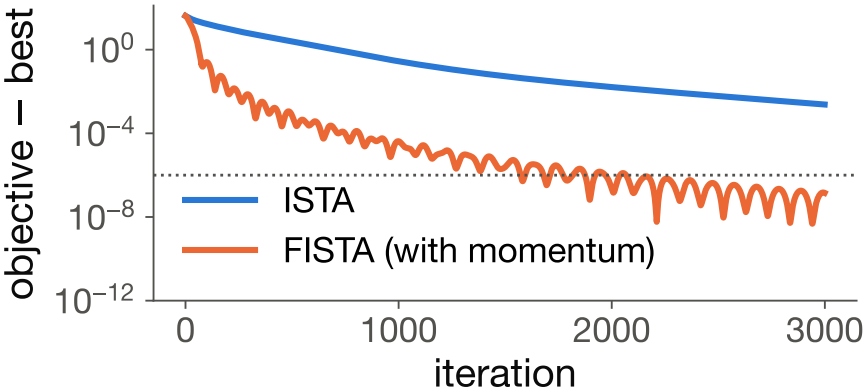
$f = \frac{1}{2}\|y - Xw\|^2 \cdot \eta = 1/L$, L: largest eigenvalue of $X^\top X$

Lasso by ISTA on the 20 features



$\lambda = 12$: 4 true kept · 13 exactly 0 · 3 tiny

FISTA: add momentum



strongly correlated features: ill-conditioned (Lecture 4)

FISTA: 10^{-6} at 1574 · ISTA: 2.4×10^{-3} at 3000

Summary: soft-thresholding and ISTA

- Let $f(w) = \frac{1}{2}\|y - Xw\|^2$, the squared error on data X, y.
- We want: $\min_w \; f(w) + \lambda\|w\|_1$, but $|w|$ is not differentiable at 0.
- Proximal operator: $S_\lambda(z) = \text{sign}(z)\max(|z| - \lambda, 0)$ minimizes $\frac{1}{2}(w - z)^2 + \lambda|w|$.
- ISTA: $z = w - \eta\nabla f(w)$, $w = S_{\eta\lambda}(z)$ entry by entry, repeat; $\eta = 1/L$, L the largest curvature.
- ISTA is slow when ill-conditioned; FISTA adds momentum

Checkpoint 2

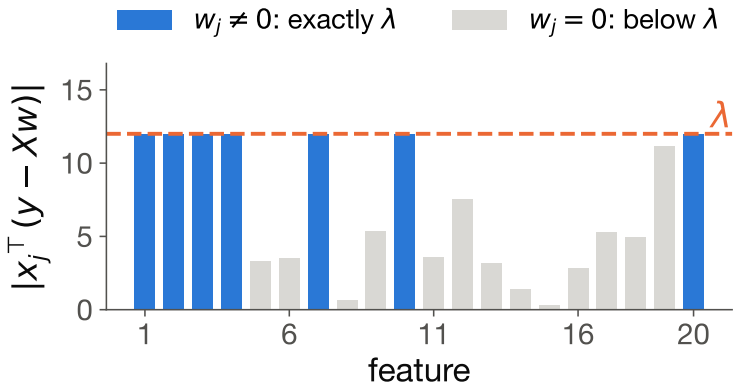
$$S_1\big(3,\ 0.5,\ -1.5\big) = ?$$

Checkpoint 2

$$S_1(3,\ 0.5,\ -1.5) = ?$$

$$(2,\ 0,\ -0.5)$$

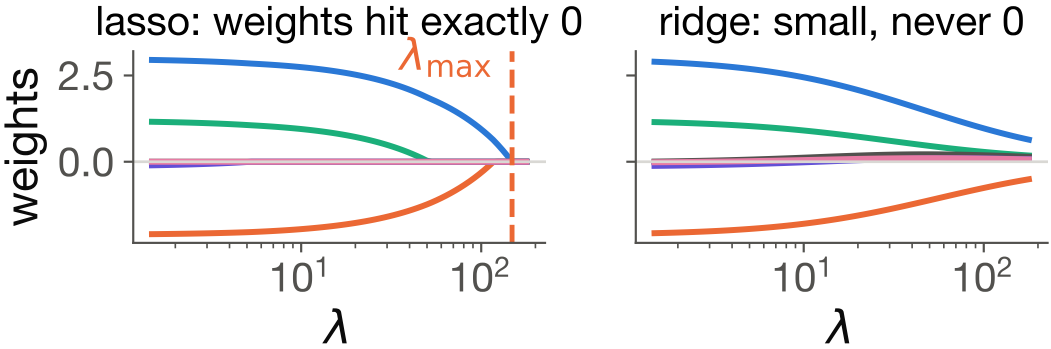
How to check a lasso answer



$$x_j^\top (y - Xw) = \lambda \operatorname{sign}(w_j) \text{ if } w_j \neq 0 \quad (x_j\text{: column } j \text{ of } X)$$

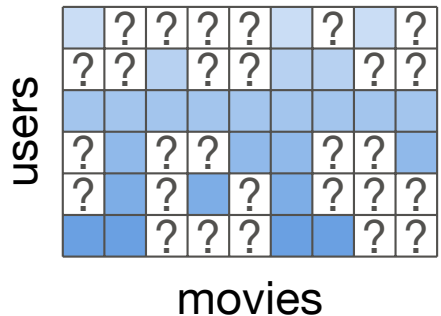
$$|x_j^\top (y - Xw)| \leq \lambda \text{ if } w_j = 0 \quad (y - Xw\text{: the residual})$$

Above $\lambda_{\max}$, every weight is zero



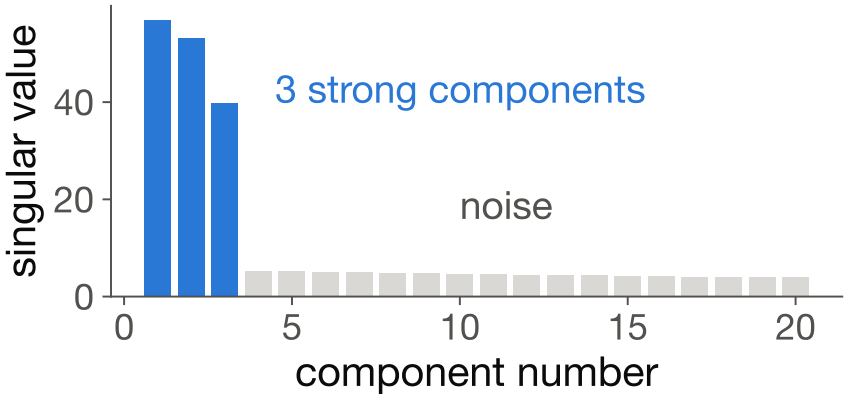
$\lambda_{\max} = \max_j |x_j^\top y|$: a natural place to start the path

Matrix completion: missing ratings



fill in the ? to recommend movies

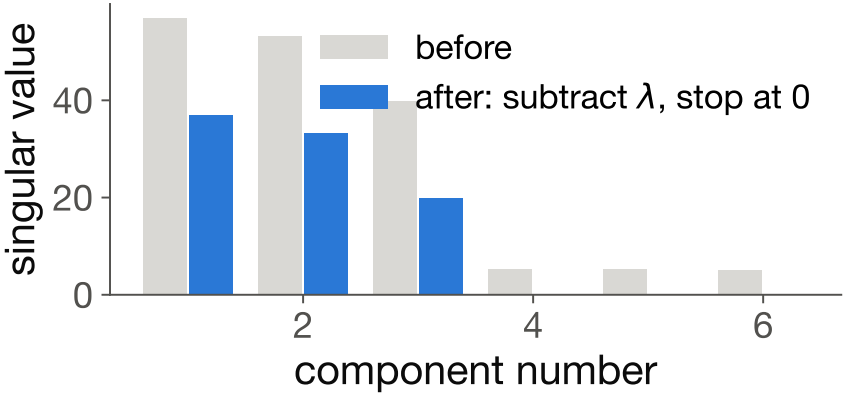
Singular values of the ratings



singular values σ_k: σ_k^2 = eigenvalues of $Z^\top Z$ (Lecture 4)

singular value: the size of each component

Shrink the singular values

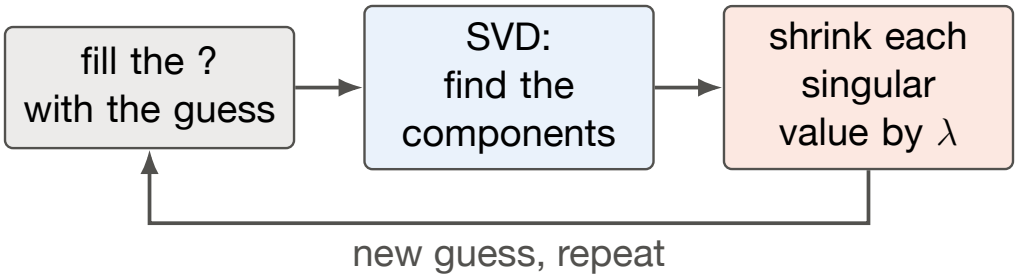


$\|Z\|_* = \sigma_1 + \sigma_2 + \ldots$ proximal operator:

soft-threshold each σ_k

here $\lambda = 20$

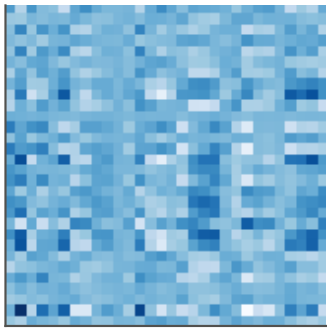
Soft-Impute



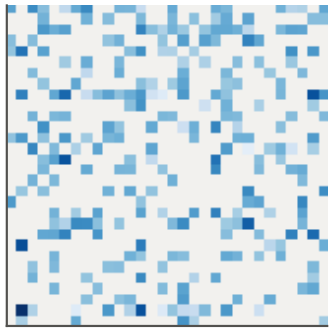
each fill step puts the observed ratings back

Soft-Impute on simulated ratings

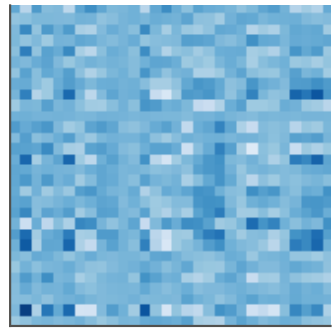
truth



observed (30%)



Soft-Impute



typical error at hidden entries, against the true ratings

Soft-Impute 0.32 vs 0.98 (fill with the average)

Summary: nuclear norm, Soft-Impute

- Let Z be a full ratings matrix, σ_k its singular values.
- Let $\|Z\|_* = \sigma_1 + \sigma_2 + \ldots$; rank = the number of $\sigma_k > 0$.
- We want: $\min_Z \ \lambda\|Z\|_* + \frac{1}{2}$ squared error on observed ratings.
- Proximal operator: $\sigma_k \to \max(\sigma_k - \lambda, 0)$; each new 0 lowers the rank.
- Soft-Impute: fill the ? with the current Z, SVD, shrink, repeat.

Checkpoint 3

singular values $5,\ 2,\ 0.5,$ $\lambda = 1$

after shrinking: ?

rank: ?

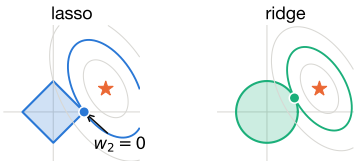
Checkpoint 3

singular values $5,\ 2,\ 0.5,$ $\lambda = 1$

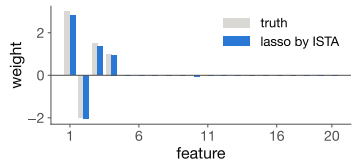
after shrinking: ? rank: ?

$4,\ 1,\ 0$ rank $3 \to 2$

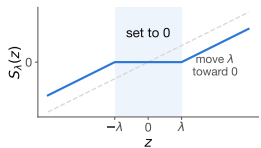
Lecture summary



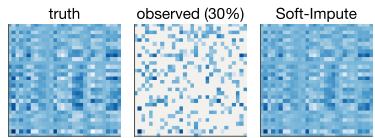
kinks give exact zeros



ISTA: gradient step, shrink

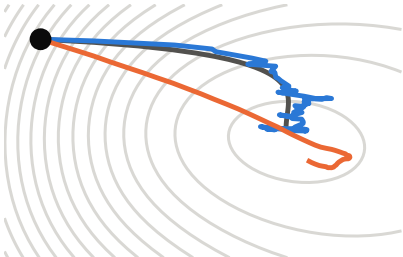


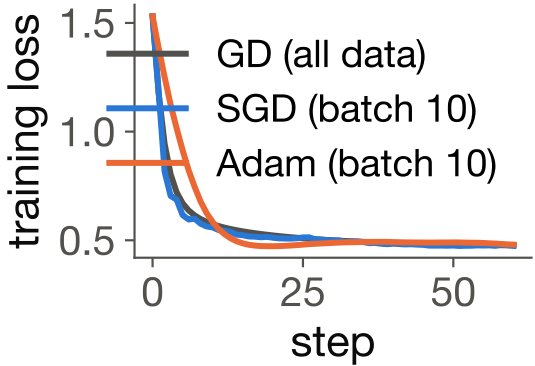
soft-thresholding



shrink singular values

Next lecture: SGD and Adam





too much data: one small batch per step

Exercise

1. $\|(1, -4, 2)\|_1 = ?$
2. $S_2\big(5,\ -1,\ -3,\ 0.5\big) = ?$
3. $f(w) = \frac{1}{2}(w-4)^2$, $\lambda = 1$, $\eta = 1$, $w_0 = 0$: one ISTA step.
4. Singular values $10,\ 3,\ 1$ and $\lambda = 2$: shrink them. What is the rank?