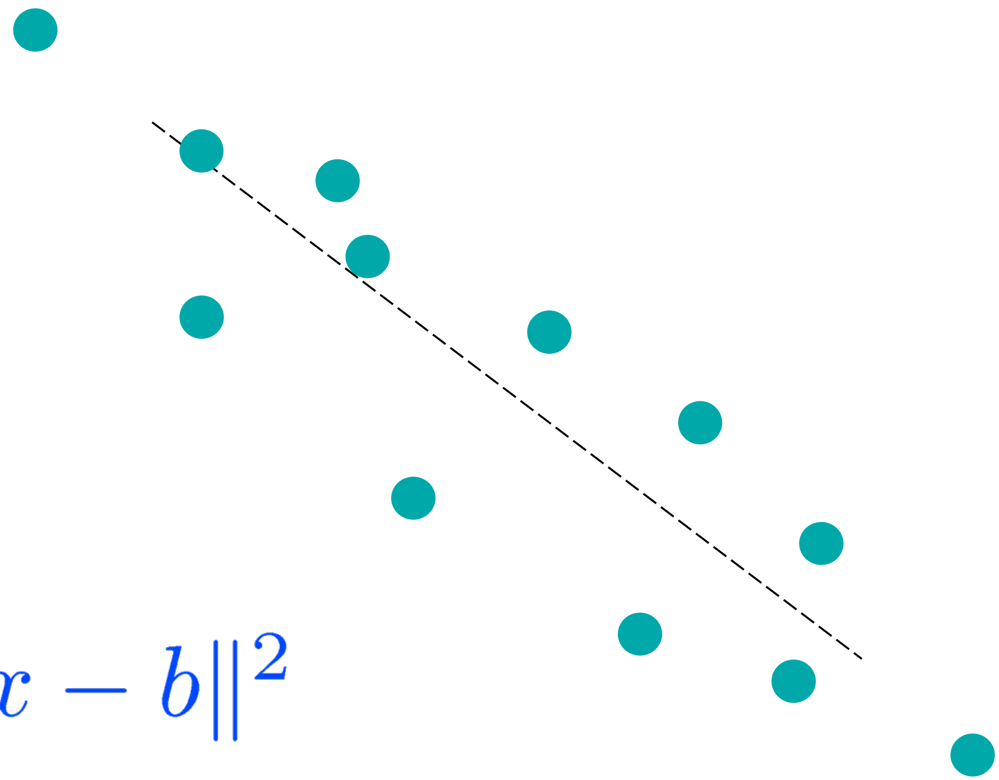
Lecture 18
Optimization approaches to Sparse Regularized Regression

Least Squares Linear Regression



$$\min \|Ax - b\|^2$$

Lasso

Primal-Dual pair of problems

$$\min \quad \frac{1}{2}||Ax-b||^2 + \lambda||x||_1$$

$$\begin{aligned} \min \quad & \frac{1}{2}x^\top A^\top A x \\ s.t. \quad & \|A^\top(Ax-b)\|_\infty \leq \lambda \end{aligned}$$

Optimality Conditions

$$\text{(i)}\quad x_i < 0, \text{ and } (A^\top(Ax-b))_i = \lambda,$$

$$\text{(ii)}\quad x_i > 0, \text{ and } (A^\top(Ax-b))_i = -\lambda,$$

$$\text{(iii)}\quad x_i = 0, \text{ and } -\lambda \leq A^\top(Ax-b)_i \leq \lambda$$

An active set approach

Optimality Conditions

$$\text{(i)}\ \ x_i < 0, \text{ and } (A^\top(Ax - b))_i = \lambda,$$

$$\text{(ii)}\ \ x_i > 0, \text{ and } (A^\top(Ax - b))_i = -\lambda,$$

$$\text{(iii)}\ \ x_i = 0, \text{ and } -\lambda \leq A^\top(Ax - b)_i \leq \lambda \text{ - relax.}$$

Given any x we partition $I = \{1, \ldots, n\}$ into I_p, I_n and I_z:

- $\forall i \in I_p\ x_i > 0$.
- $\forall i \in I_n\ x_i < 0$.
- $\forall i \in I_z\ x_i = 0$.

Active set approach

Given a partition (x_p, x_n, x_z), $x_z = 0$.

$$\min \quad \frac{1}{2}||A_p x_p + A_n x_n - b||^2 + \lambda \|(x_p, x_n)\|_1$$

Check $\|A^\top (A_{(p,n)} x_{(p,n)} - b)\|_\infty \leq \lambda$

Active set approach

$$\min \quad \frac{1}{2}||A_p x_p + A_n x_n - b||^2 + \lambda \sum_{i \in I_p} x_i - \lambda \sum_{i \in I_n} x_i$$

- Get a solution (x_p^*, x_n^*) by solving a system of linear equations
- Check if $x_p^* > 0$, and $x_n^* < 0$, if yes, continue...
- If not, find

$$i^* = \operatorname{argmin}\{\min_{i \in I_p:\, x_i^* < 0} x_i/(x_i - x_i^*), \min_{i \in I_n:\, x_n^* > 0} -x_i/(x_i^* - x_i)\}$$

- Move i^* to I_z, update I_p and I_n
- Repeat the step

Checking optimality, choosing next nonzero element

$$\text{Check } \|A^\top(A_{(p,n)}x_{(p,n)} - b)\|_\infty \leq \lambda$$

- Given solution $(x_p, x_n, 0)$ we know

$$({A_p}^\top(A_px_p + A_nx_n - b))_i = \lambda,$$
$$({A_n}^\top(A_px_p + A_nx_n - b))_i = -\lambda,$$

- Check if

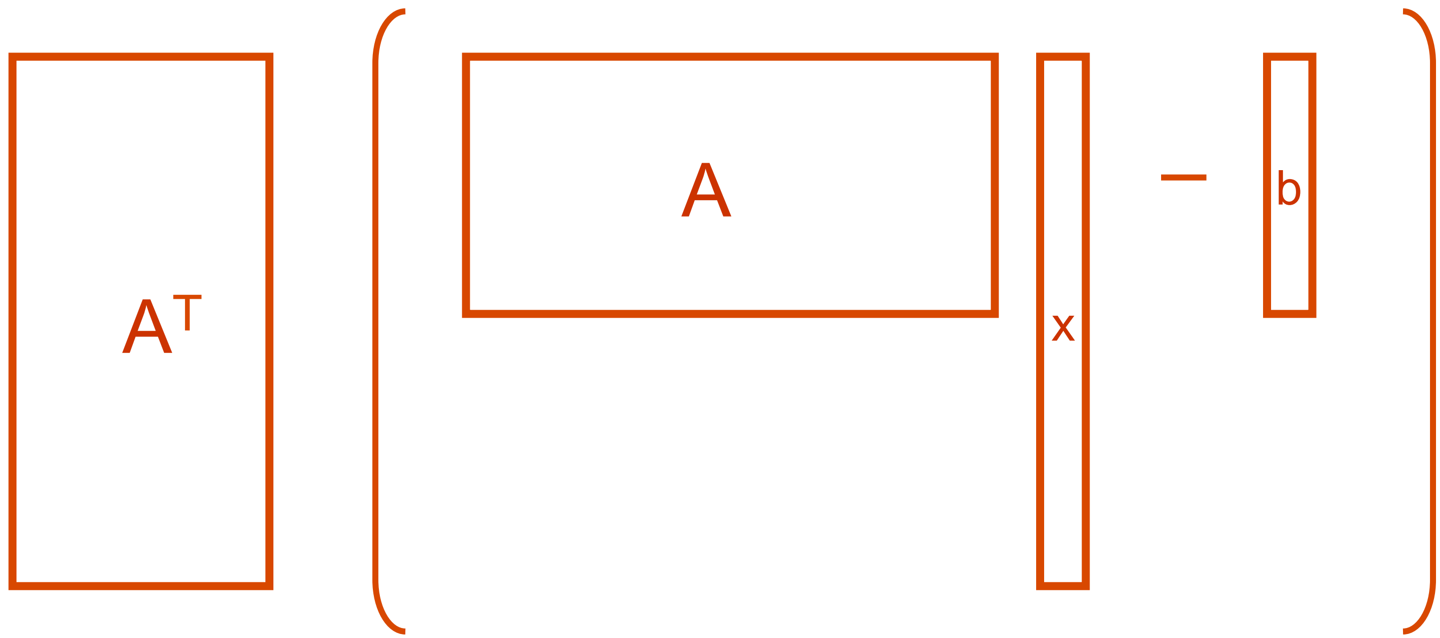
$$-\lambda \leq ({A_z}^\top(A_px_p + A_nx_n - b))_i \leq \lambda$$

  if true, then optimal solution is reached, otherwise...

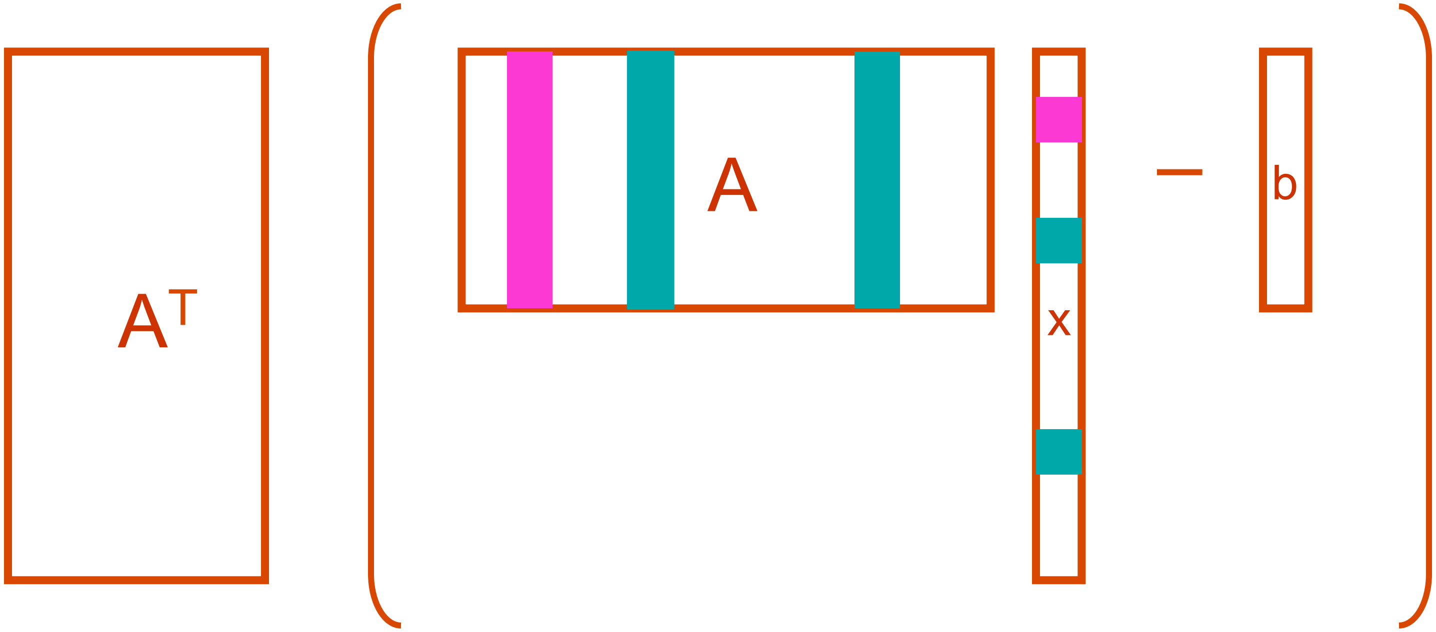
- Choose $i^* = \text{argmax}_{i\in I_z}\{-\min({A_z}^\top(A_px_p + A_nx_n - b))$
  $\max({A_z}^\top(A_px_p + A_nx_n - b))\}$

- Move i^* into I_p or I_n according to the sign of $({A_z}^\top(A_px_p + A_nx_n - b))_{i^*}$, update I_z.

Least angle regression



Least angle regression

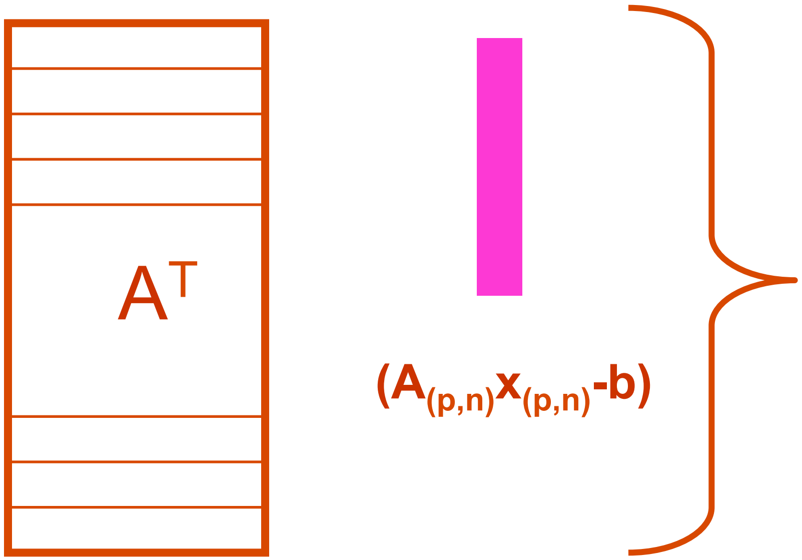


Choose I_p and I_n compute $x_{(n,p)}$ from

(i) $i \in I_n$, and $(A^\top(A_p x_p + A_n x_n - b))_i = \lambda$,

(ii) $i \in I_p$, and $(A^\top(A_p x_p + A_n x_n - b))_i = -\lambda$.

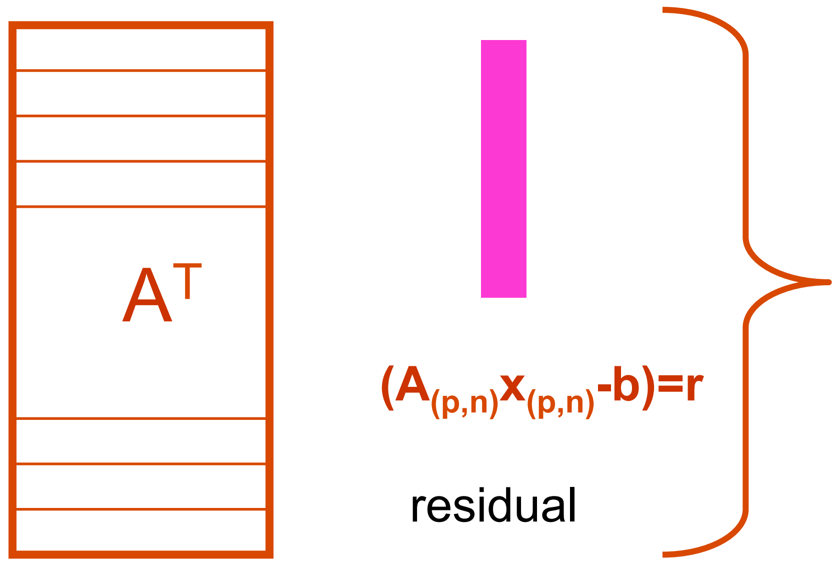
Least angle regression



Choose index i with the largest $|A^{\top}(A_{(p,n)}x_{(p,n)} - b)|_i$

$$i \in I_z, \text{ check } -\lambda \le (A^{\top}(A_{(p,n)}x_{(p,n)} - b))_i \le \lambda$$

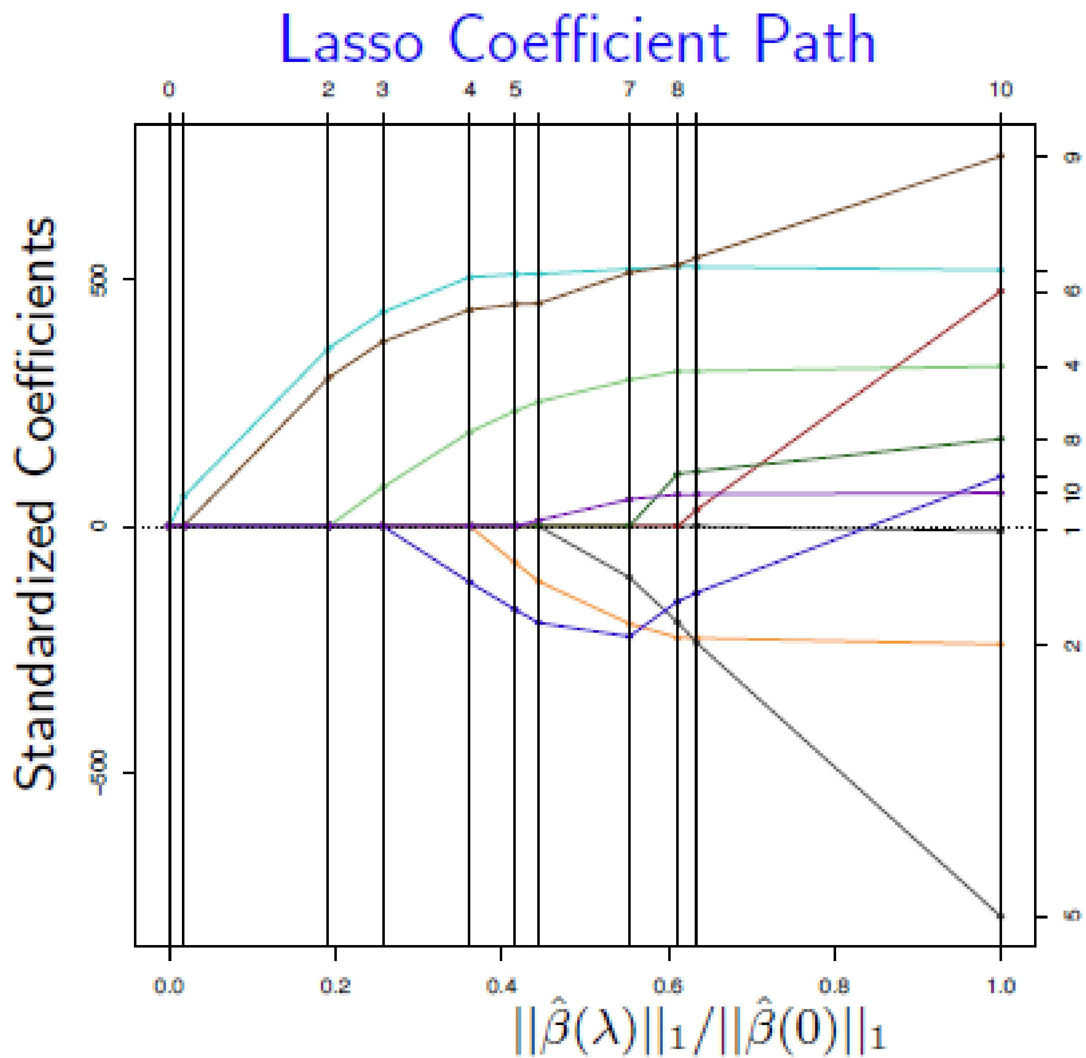
Least angle regression



Choose index i with the largest $|(A^\top(A_{(p,n)}x_{(p,n)} - b))_i|$

The largest $|{A_i}^\top r|$ is given by the column of of A which makes the least angle with r or with $-r$ - the largest positive or negative correlation.

If all angles are “big” (defined by λ) or r is small then we are done!



Lasso: $\hat{\beta}(\lambda) = \text{argmin}_\beta \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda||\beta||_1$

Computing regularization path

Let us start with a very large λ and $I_z = \{1, \ldots, n\}$, $|(A^\top b)_i| \leq \lambda$

- Reduce λ until for some $i^* \in I_z$ $\lambda = |(A^\top b)|_{i^*}$ occurs.
- Move i^* into I_p or I_n according to the sign of $({A_z}^\top (A_p x_p + A_n x_n - b))_{i^*}$, update I_z.
- Keep reducing λ until either $\lambda = |(A^\top b)|_{i^*}$ for some $i^* \in I_z$ or for solution (x_p, x_n) which satisfies

$$
\begin{aligned}
({A_p}^\top (A_p x_p + A_n x_n - b))_i &= \lambda,\ i \in I_n \\
({A_n}^\top (A_p x_p + A_n x_n - b))_i &= -\lambda,\ I \in I_p
\end{aligned}
$$

  one of the components hits zero.
- Update I_z, I_p and I_n and proceed reducing λ.

Per iteration cost

$$\text{Given a partition } (x_p, x_n, 0),\ |I_p \cup I_n| = k,$$

$$\min \quad \frac{1}{2}||A_{(p,n)}x_{(p,n)} - b||^2 + \lambda \sum_{i \in I_p} x_i - \lambda \sum_{i \in I_n} x_i$$

Update factorization of $A_{(p,n)}{}^{T}A_{(p.n)}$ at each step - O(mk)

Memory – $O(k^2)$

$$\text{Check } \|A^{\top}(A_{(p,n)}x_{(p,n)} - b)\|_{\infty} \leq \lambda$$

Compute $A^T(A_{(p,n)}x_{(p,n)})$: O(nm) (improved by “sifting”)

Can be too costly to compute and to store.

Coordinate descent

Coordinate descent

Choose one variable x_i and column A_i.
Let $\bar{x}$ and $\bar{A}$ correspond to the fixed part

$$\min_{x_i} \quad \frac{1}{2}||A_i x_i + \bar{A}\bar{x} - b||^2 + \lambda|x_i|$$

Soft-thresholding operator

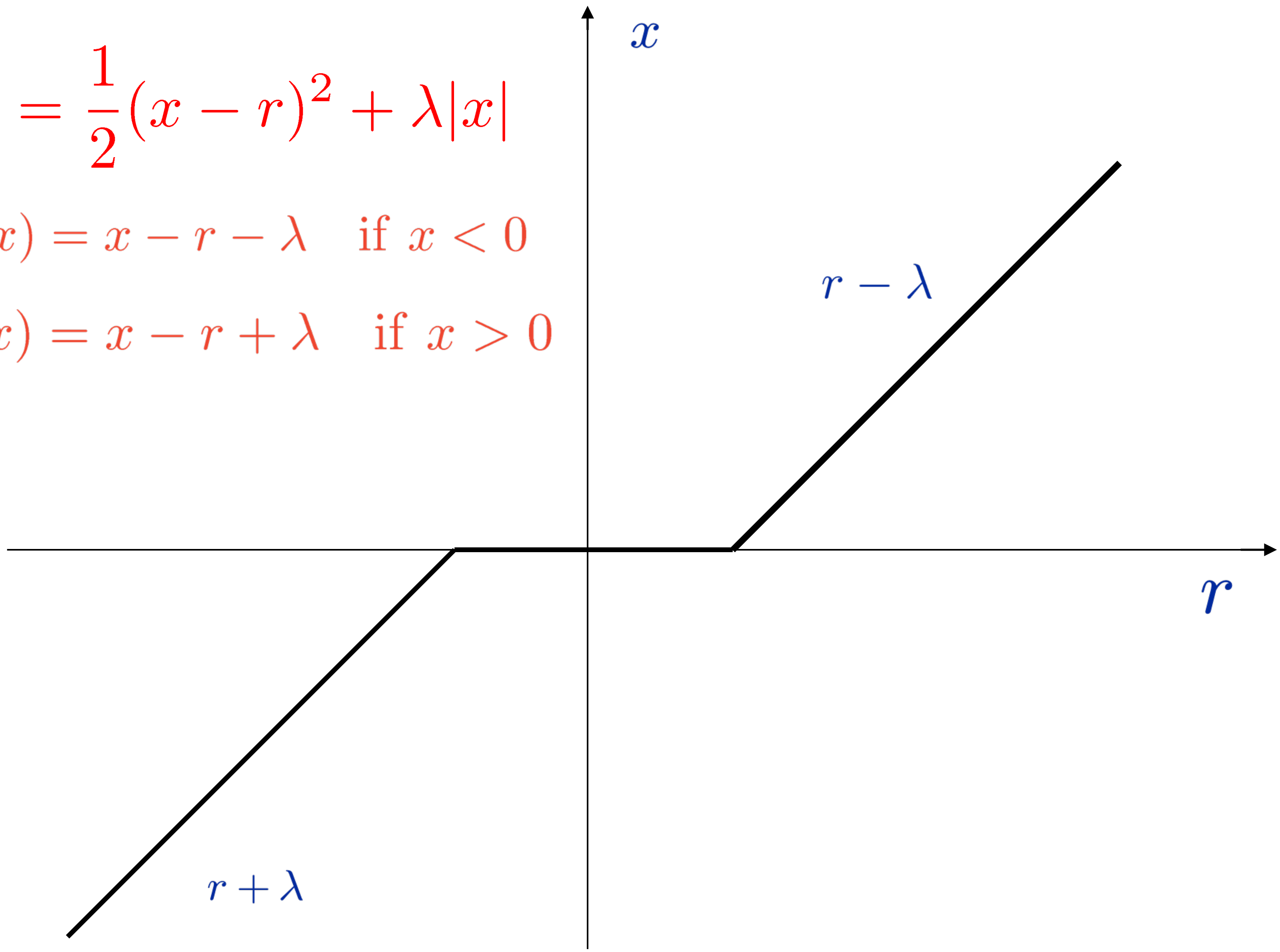
$$\min_{x_i} \frac{1}{2}(x_i - r)^2 + \lambda|x| \rightarrow x_i = \begin{cases} r - \lambda & \text{if } r > \lambda \\ 0 & \text{if } -\lambda \leq r \leq \lambda \\ r + \lambda & \text{if } r < -\lambda \end{cases}$$

$$r = -{A_i}^\top(\bar{A}\bar{x} - b)/||A_i||^2, \ \lambda \rightarrow \lambda/||A_i||^2$$

$$f(x) = \frac{1}{2}(x - r)^2 + \lambda|x|$$

$$\nabla_x f(x) = x - r - \lambda \quad \text{if } x < 0$$

$$\nabla_x f(x) = x - r + \lambda \quad \text{if } x > 0$$



Given the scaled gradient

$$r: \;\; r_i = {A_i}^\top(\bar{A}\bar{x} - b)/||A_i||,$$

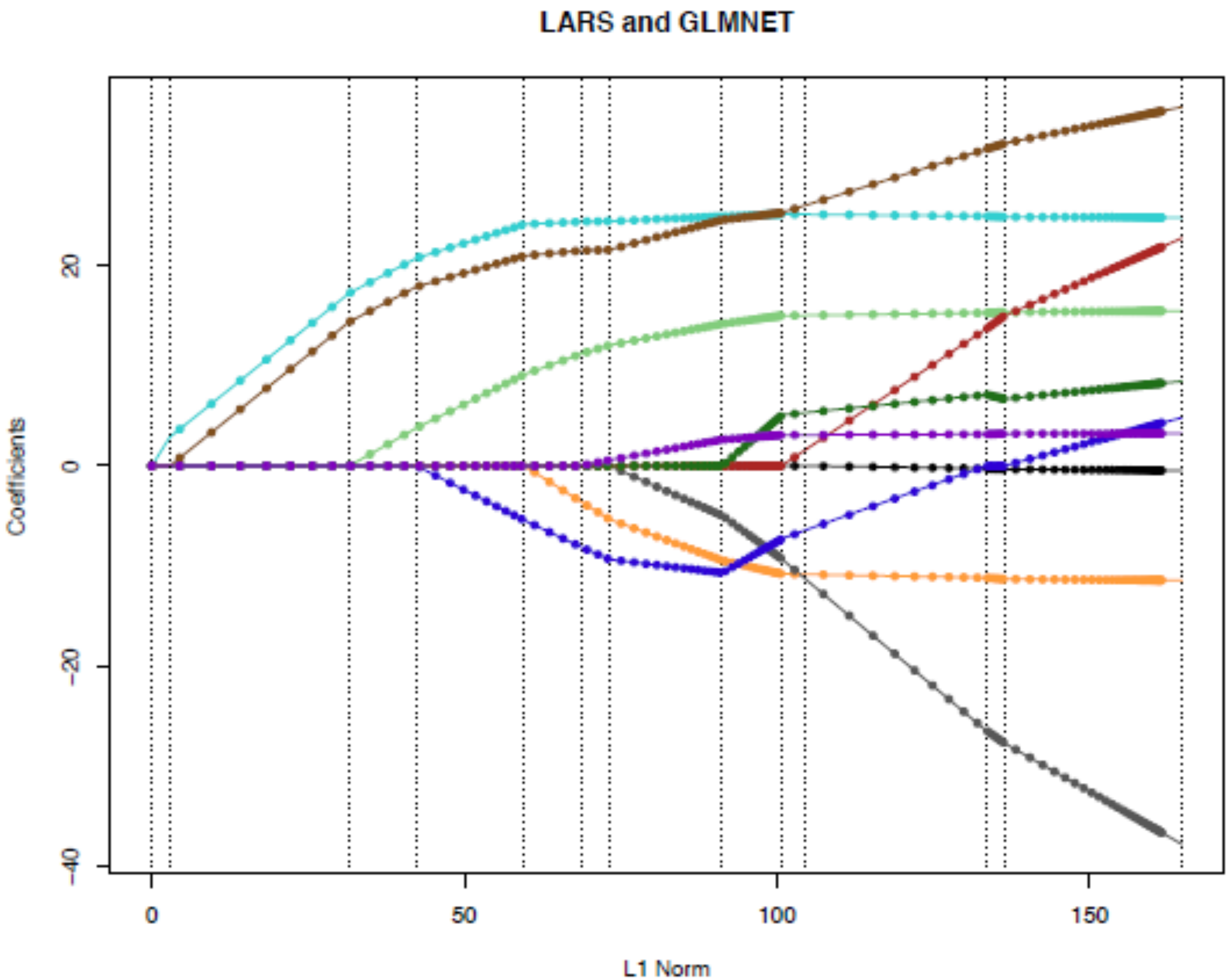
Can choose coordinate to update by:

- Simply cycle through all coordinates
- Update all at once
- Choose the one with largest gradient component
- Choose the one with largest obj. function improvement
- Choose coordinate at random

Coordinate Descent

- Solve the lasso problem by coordinate descent: optimize each parameter separately, holding all the others fixed. Updates are trivial. Cycle around till coefficients stabilize.
- Do this on a grid of λ values, from λ_{max} down to λ_{min} (uniform on log scale), using warms starts.
- Can do this with a variety of loss functions and additive penalties.

Coordinate descent achieves dramatic speedups over all competitors, by factors of 10, 100 and more.


LARS and GLMNET
Coefficients
L1 Norm
20
0
−20
−40
0
50
100
150

First order methods

First-order proximal gradient methods

- Consider: $\boxed{\min_x f(x)}$

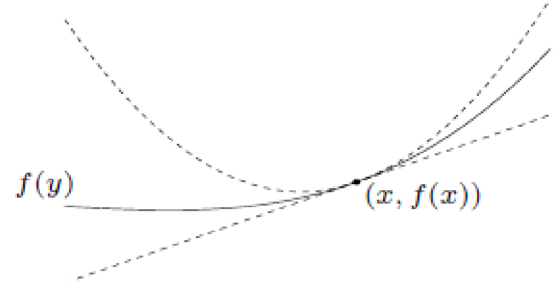
$$|\nabla f(x) - \nabla f(y)| \leq L||x-y||$$

- Linear lower approximation

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x)$$

- Quadratic upper approximation

$$f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2\mu}||y-x||^2 = Q_{f,\mu}(x,y)$$



$$f(y) \leq f(x) + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 = Q_{f,\mu}(x,y)$$

First-order proximal gradient method

$$\boxed{\min_x f(x)}$$

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \mathrm{argmin}_y Q_{f,\mu}(x^k, y)$$

$$\Updownarrow$$

$$x^{k+1} = x^k - \mu \nabla f(x^k)$$

- If $\mu \leq$ *1/L* then

$$f(x^{k+1}) \leq f(x^k) + \frac{1}{2\mu}||x^k - \mu\nabla f(x^k)^\top - x^{k+1}||^2 = Q_{f,\mu}(x^k, x^{k+1})$$

Accelerated first-order method

Nesterov, ’83, ‘00s,

Beck&Teboulle ‘09

$$\boxed{\min_x f(x)}$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = y^k - \mu \nabla f(y^k)$$

$$y^{k+1} := \quad x^k + \tfrac{k-1}{k+2}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then

$$f(x^k) - f(x^*) \leq \frac{L\|x^0 - x^*\|}{2k^2}$$

Complexity of accelerated first-order method

Nesterov, ’ 83, ‘00s,

Beck&Teboulle ‘09

$$\boxed{\min_x f(x)}$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = y^k - \mu \nabla f(y^k)$$

$$y^{k+1} := \quad x^k + \tfrac{k-1}{k+2}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then in $O(\sqrt{\frac{L\|x^0 - x^*\|}{\epsilon}})$ iterations finds solution

$$\bar{x} : \; f(\bar{x}) \leq f(x^*) + \epsilon$$

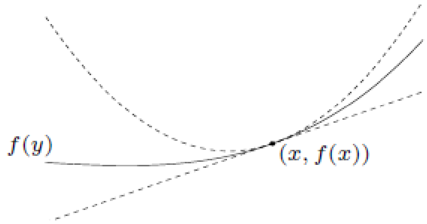
This method is optimal if only gradient information is used.

Prox method with nonsmooth term

- Consider: $\boxed{\min_x F(x) = f(x) + g(x)}$

$$|\nabla f(x) - \nabla f(y)| \le L||x - y||$$

- Quadratic upper approximation



$$f(y) + g(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y) = Q_{f,\mu}(x, y)$$

$$F(y) \le f(x) + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 + g(y) = Q_{f,\mu}(x, y)$$

Assume that *g(y)* is such that the above function is easy to optimize over *y*

First-order method for nonsmooth functions

$$\boxed{\min_x F(x) = f(x) + g(x)}$$

- Minimize quadratic upper relaxation on each iteration

$$x^{k+1} = \mathrm{argmin}_y Q_f(x^k, y) = f(x^k) + \frac{1}{2t}||x^k - \mu\nabla f(x^k)^\top - y||^2 + g(y)$$

- If $\mu \leq$ *1/L* then in *O(1/ϵ)* iterations finds solution

$$\bar{x} : \ F(\bar{x}) \leq F(x^*) + \epsilon$$

Beck&Teboulle, Tseng,
Auslender&Teboulle, 2008

Fast-first order methods

$$\boxed{\min_x F(x) = f(x) + g(x)}$$

- Minimize a upper approximation at an intermediate point.

$$x^{k+1} = \mathrm{argmin}_y Q_f(y^k, y)$$

$$y^{k+1} = x^k + \frac{k-1}{k+2}(x^{k+1} - x^k)$$

- If $\mu \leq$ *1/L* then in $O(1/\sqrt{\epsilon})$ iterations finds solution

$$\bar{x} : \; F(\bar{x}) \leq F(x^*) + \epsilon$$

Beck&Teboulle, Tseng, 2008

g(y) in sparse regression

Example 1 (Lasso)

$$\boxed{\min_x f(x) + ||x||_1}$$

- Minimize upper approximation function $Q_f(x,y)$ on each iteration

$$\min_y f(x^k) + \frac{1}{2t}||x^k - t\nabla f(x^k)^\top - y||^2 + \|y\|_1$$

$$\Updownarrow$$

$$\sum_i \min_{y_i} \left[\frac{1}{2t}(y_i - r_i)^2 + |y_i|\right]$$

$$\Updownarrow$$

Closed form solution! $O(n)$ effort

$$\min_{y_i} \frac{1}{2}(y_i - r_i)^2 + t|y_i| \rightarrow y_i^* = \begin{cases} r_i - t & \text{if } r_i > t \\ 0 & \text{if } -\lambda \leq r_i \leq t \\ r_i + t & \text{if } r_i < -t \end{cases}$$

Gradient method for Lasso

$$\nabla f(x) = A^\top (Ax - b)$$

$$x^{k+1} = \min_y (Ax^k - b)^\top A(x^k - y) + \frac{1}{2t}||x^k - y||^2 + \lambda\|y\|_1$$

$$x^{k+1} = \min_y \frac{1}{2t}||(x^k - tA^\top (Ax^k - b)) - y||^2 + \lambda\|y\|_1$$

⬇

2 matrix/vector multiplications + shrinkage operator per iteration

O(1/ϵ) iteration bound

Sparse logistic regression

$$f(w,\beta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i(w^\top x_i + \beta)))$$

$$\min_{w,\beta} f(w,\beta) + \lambda\|w\|_1$$

$$w^{k+1} = \min_{v} \frac{1}{2\mu}||(w^k - \mu\nabla_w f(w^k,\beta) - v||^2 + \lambda\|v\|_1$$

$$\Downarrow$$

A gradient computation + shrinkage operator per iteration

$O(1/\epsilon)$ iteration bound

Example 2 (Group Lasso)

$$\min_x f(x) + \sum_i ||x_i||, \;\; x_i \in \mathrm{R}^{n_i}$$

- Very similar to the previous case, but with ||.|| instead of |.|

$$\sum_i \min_{y_i \in \mathrm{R}^{n_i}} \left[\frac{1}{2t}(y_i - r_i)^2 + \|y_i\| \right]$$

$$\Updownarrow$$

$$y_i^* = \frac{r_i}{\|r_i\|} \max(0, \|r_i\| - \mu)$$

Closed form solution!
O(n) effort

SIMPLIFIED ACTIVE SET
(EXTRA SLIDES)

Optimality Conditions

$$\text{(i)}\quad x_i < 0, \text{ and } (A^\top(Ax-b))_i = \lambda,$$

$$\text{(ii)}\quad x_i > 0, \text{ and } (A^\top(Ax-b))_i = -\lambda,$$

$$\text{(iii)}\quad x_i = 0, \text{ and } -\lambda \leq A^\top(Ax-b)_i \leq \lambda \text{ - relax.}$$

Given any x we partition $I = \{1, \ldots, n\}$ into B and N:

- $\forall i \in B \; x_i \neq 0$.
- $\forall i \in N \; x_i = 0$.

Active set approach

Given a partition (x_B, x_N), $x_N = 0$.

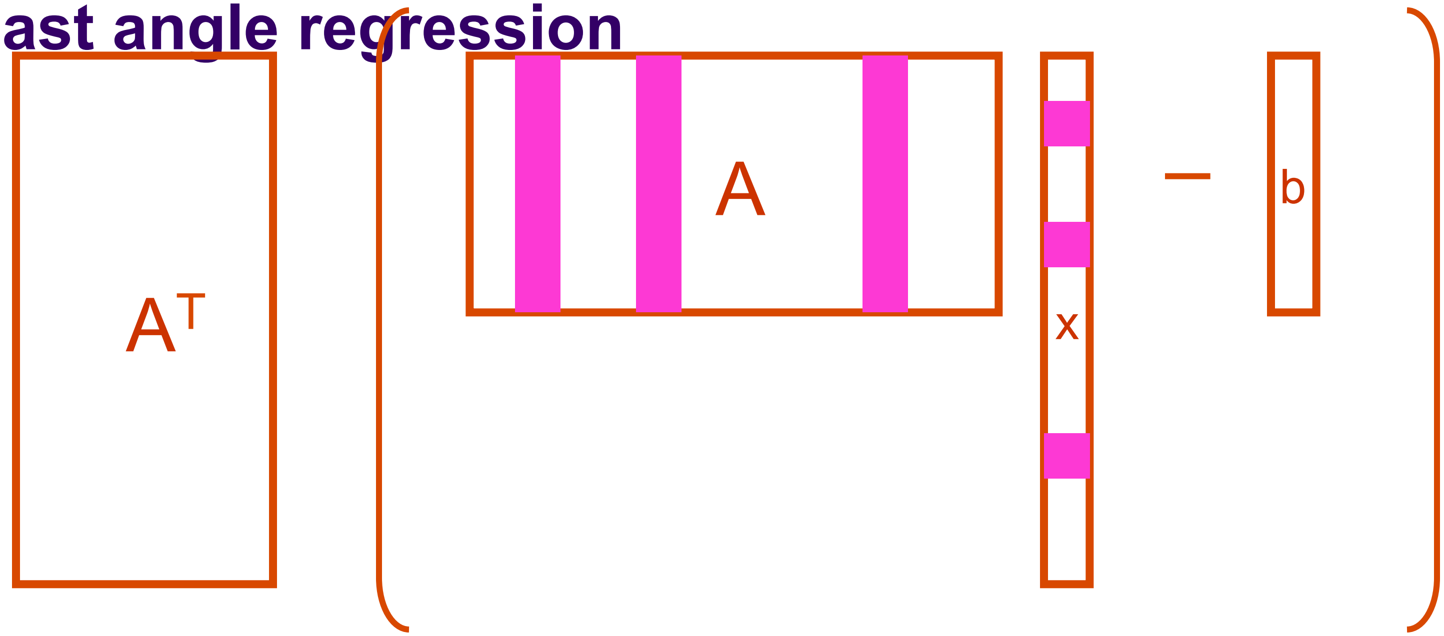
$$\min \quad \frac{1}{2}||A_B x_B - b||^2 + \lambda \|x_B\|_1$$

Check $\|A^\top (A_B x_B - b)\|_\infty \leq \lambda$

Least angle regression

$$A^T \left(A \, x - b \right)$$

Least angle regression

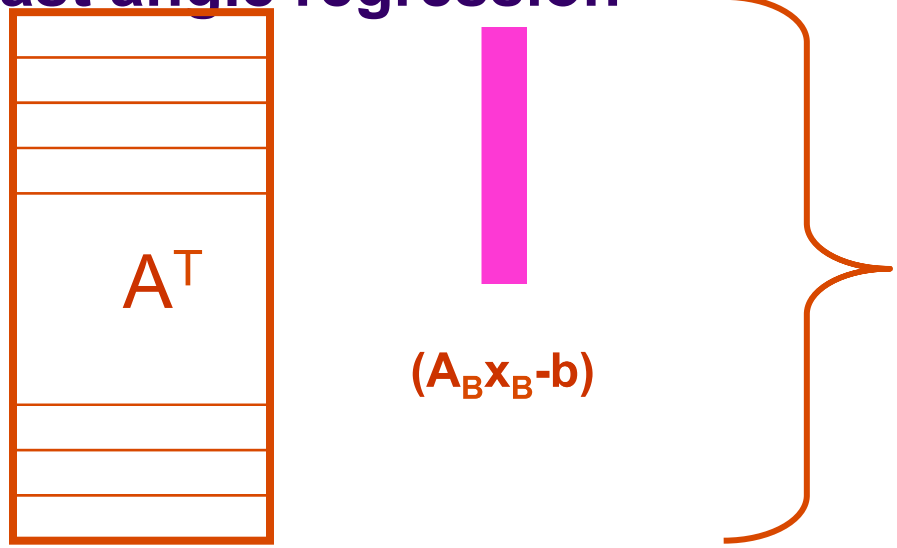


Choose B and compute x_B from

(i) $x_i < 0$, and $(A^\top(A_B x_B - b))_i = \lambda$,

(ii) $x_i > 0$, and $(A^\top(A_B x_B - b))_i = -\lambda$.

Least angle regression



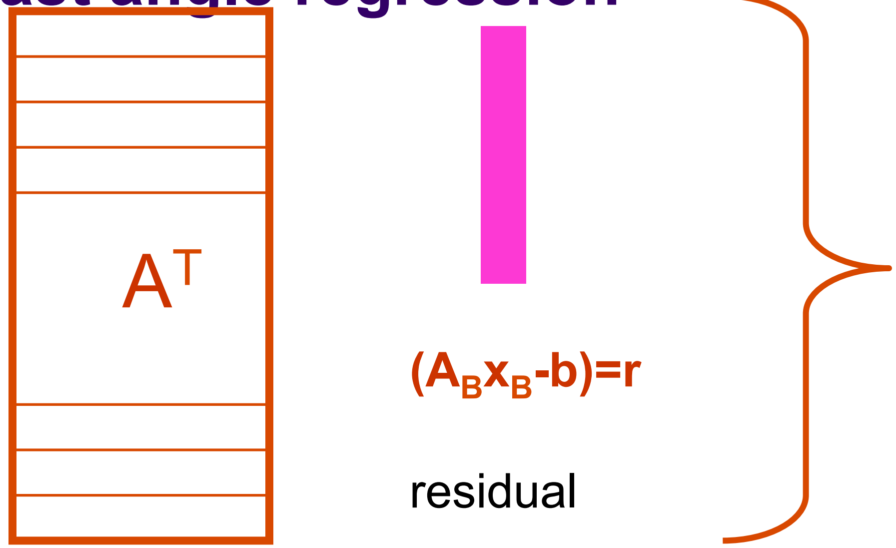
Choose index i with the largest $|A^\top(A_B x_B - b)|_i$

$$\text{(i)} \quad x_i < 0, \text{ and } (A^\top(A_B x_B - b))_i = \lambda,$$

$$\text{(ii)} \quad x_i > 0, \text{ and } (A^\top(A_B x_B - b))_i = -\lambda,$$

$$\text{(iii)} \quad x_i = 0, \text{ and } -\lambda \le A^\top(A_B x_B - b)_i \le \lambda \text{ - relax.}$$

Least angle regression



Choose index i with the largest $|A^\top(A_B x_B - b)|_i$

The largest $|{A_i}^\top r|$ is given by the column of A which makes the least angle with r or with $-r$ - the largest positive or negative correlation.

Per iteration cost

Given a partition (x_B, x_N), $x_N = 0$.

$$\min \quad \frac{1}{2}||A_B x_B - b||^2 + \lambda \|x_B\|_1$$

Update factorization of $A_B^T A_B$: $O(mk)$ if $A_B \in R^{m \times k}$

Check $\|A^\top (A_B x_B - b)\|_\infty \leq \lambda$

Compute $A^T(A_B x_B)$: $O(nm)$ (can be improved in practice)

Can be too costly to compute and to store. We'll now see how to avoid any matrix factorizations