Complementary material

Sparse representation

L1-norm solution

L1-norm for sparsity

$$\min_{c} \|c\|_0, \text{ s.t. } Ac = x. \quad \textbf{P1}$$

- The problem is NP-hard. Greedy approaches such as MP and OMP have been introduced.
- Another solution is to relax the problem P1 to an $L1$-norm minimization problem.

$$\min_{c} \|c\|_1, \text{ s.t. } Ac = x., \quad \textbf{PL1}$$

where $\|c\|_1$ is the one-norm of c, i.e., sum of the absolute values, $\|c\|_1 = |c_1| + |c_2| + \cdots + |c_m|$.

L-1 norm relaxation

- Algorithms solving Problem PL1 is called the **basis pursuit** method. In Statistics, it is called LASSO.
- We can apply an optimization algorithm called **coordinate descent** to solve the LASSO problem.

Coordinate descent method for the L1-norm sparse problem

- $L1$-norm minimization problem.

$$\min_{c} \|c\|_1, \text{ s.t. } Ac = x., \quad \textbf{PL1}$$

- Consider a related form of the problem regrading noises:

$$\min_{c} \|Ac - x\|^2 + \lambda \|c\|_1, \quad \textbf{PL2}$$

where λ is a positive parameter for the L1 regularization term, $\|c\|_1$. The larger is λ, the stronger L1-norm constraint (i.e., more sparsity) is imposed.

Normalization vs Constraint

- Problem PL2 is highly related to the L1-constraint version below, but PL2 is more popular because it is an un-constrained optimization problem.

$$\min_{c} \|x - Ac\|^2,$$
$$\text{s.t. } \|c\|_1 \leq \epsilon\,.$$

Constrained optimization

$$\min_{c} \|Ac - x\|^2 + \lambda \|c\|_1$$

PL2 is unconstrained optimization

Coordinate Descent for Solving PL2

- We use the materials from the following two slides to introduce the approach.
  - **Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725**
    - **https://www.cs.cmu.edu/~ggordon/10725-F12/slides/25-coord-desc.pdf**
  - **useR! 2009 Trevor Hastie, Stanford Statistics**
    - **https://web.stanford.edu/~hastie/TALKS/glmnet.pdf**

Review of Coordinate Descent

From **Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725**

Coordinate-wise minimization

We've seen (and will continue to see) some pretty sophisticated methods. Today, we'll see an extremely **simple** technique that is surprisingly efficient and scalable

Focus is on **coordinate-wise minimization**

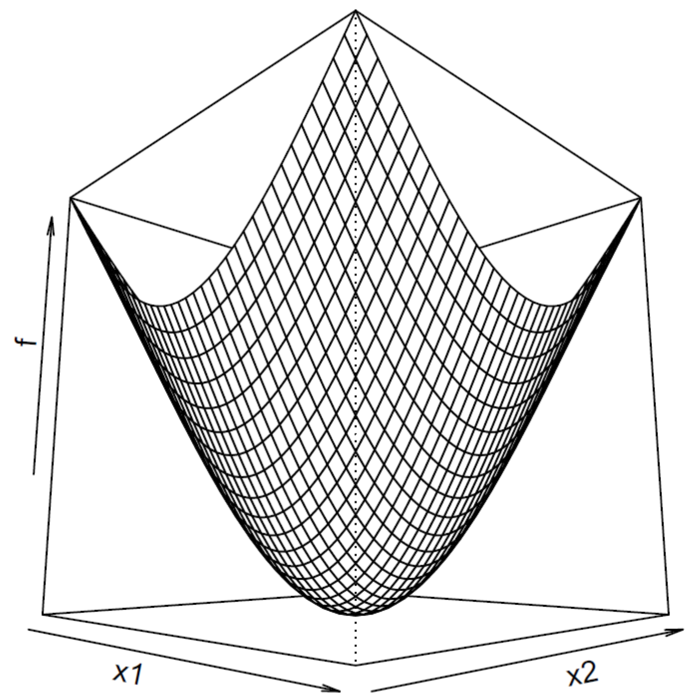
Q: Given convex, differentiable $f : \mathbb{R}^n \to \mathbb{R}$, if we are at a point x such that $f(x)$ is minimized along each coordinate axis, *have we found a global minimizer?* A: Yes!

I.e., does $f(x + d \cdot e_i) \geq f(x)$ for all $d, i \;\Rightarrow\; f(x) = \min_z f(z)$?

(Here $e_i = (0, \ldots, 1, \ldots 0) \in \mathbb{R}^n$, the ith standard basis vector)

Review of Coordinate Descent

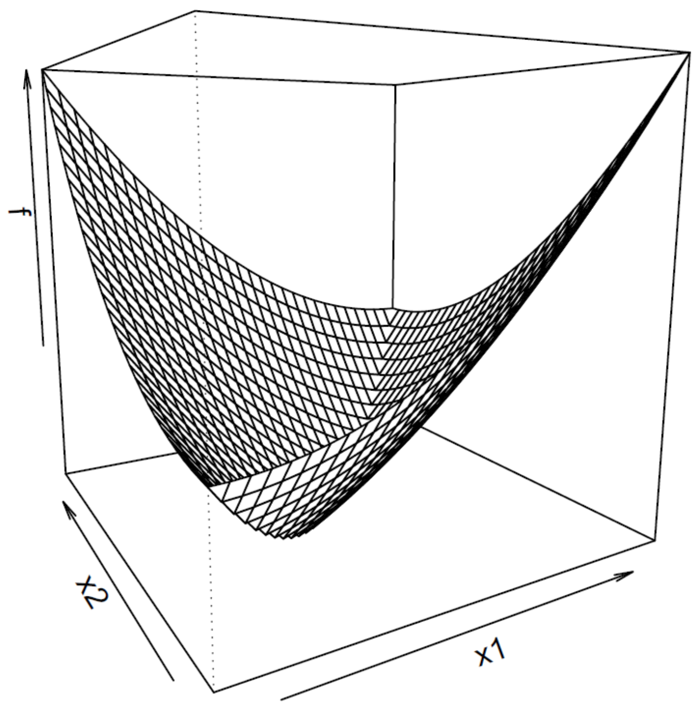
From **Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725**

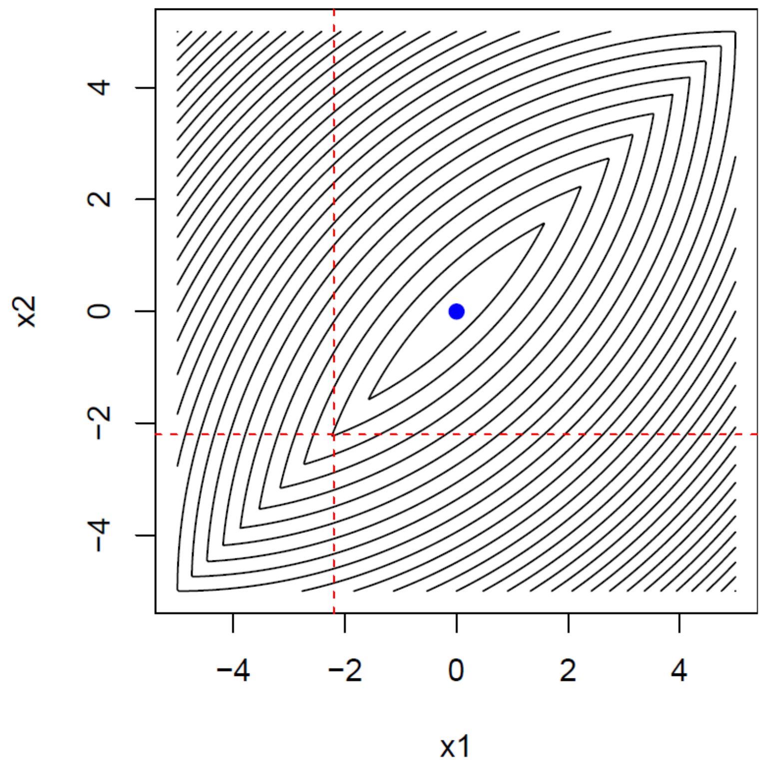


Q: Same question, but for f convex (not differentiable) ... ?

Review of Coordinate Descent

From **Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725**





A: No! Look at the above counterexample

Q: Same question again, but now $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$, with g convex, differentiable and each h_i convex ... ? (Non-smooth part here called **separable**)

A: Yes!

Review of Coordinate Descent

From **Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725**

Lasso regression

Here, y and x are our notations x and c in PL2, respectively

Consider the lasso problem

$$f(x) = \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

Note that the non-smooth part is separable: $\|x\|_1 = \sum_{i=1}^{p} |x_i|$

Minimizing over x_i, with x_j, $j \neq i$ fixed:

$$0 = A_i^T A_i x_i + A_i^T (A_{-i} x_{-i} - y) + \lambda s_i$$

where $s_i \in \partial|x_i|$. Solution is given by soft-thresholding

$$x_i = S_{\lambda/\|A_i\|^2}\left(\frac{A_i^T(y - A_{-i}x_{-i})}{A_i^T A_i}\right)$$

$S_a(\cdot)$ denotes the soft-thresholding function,
$S_a(x) = sign(x)\max(|x| - a, 0)$

Repeat this for $i = 1, 2, \ldots p, 1, 2, \ldots$

Coordinate descent for the lasso

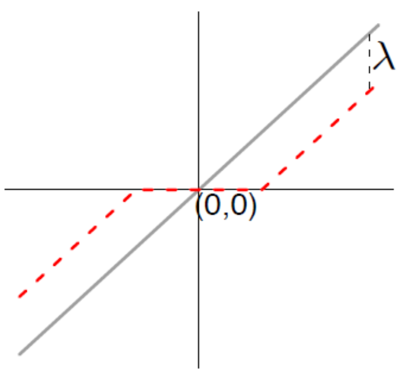
$$\min_\beta \frac{1}{2N}\sum_{i=1}^N (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Suppose the p predictors and response are standardized to have mean zero and variance 1. Initialize all the $\beta_j = 0$.

Cycle over $j = 1, 2, \ldots, p, 1, 2, \ldots$ till convergence:

- Compute the partial residuals $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$.
- Compute the simple least squares coefficient of these residuals on jth predictor: $\beta_j^* = \frac{1}{N}\sum_{i=1}^N x_{ij} r_{ij}$
- Update β_j by *soft-thresholding*:

$$\boxed{\begin{aligned} \beta_j &\leftarrow S(\beta_j^*, \lambda) \\ &= \operatorname{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+ \end{aligned}}$$



Guarantees of L1-norm Optimization for Sparse Solutions

- We use the materials summarized from Wotao Yin's slides
  - Wotao Yin: Sparse Optimization Lecture: Sparse Recovery Guarantees
    - https://www.ise.ncsu.edu/fuzzy-neural/wp-content/uploads/sites/9/2020/08/SparseRecoveryGuarantees.pdf

Examples of guarantees

Theorem (Donoho and Elad [2003], Gribonval and Nielsen [2003])

For $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ has full rank, if $\mathbf{x}$ satisfies $\|\mathbf{x}\|_0 \le \frac{1}{2}(1 + \mu(\mathbf{A})^{-1})$, then ℓ_1-minimization recovers this $\mathbf{x}$.

Recall that in OMP we have the property $\mu < \frac{1}{2k-1}$.

Same condition!

Note that the symbols m and n are different from ours (exchanged).

Theorem (Candes and Tao [2005])

If $\mathbf{x}$ is k-sparse and $\mathbf{A}$ satisfies the RIP-based condition $\delta_{2k} + \delta_{3k} < 1$, then $\mathbf{x}$ is the ℓ_1-minimizer.

Theorem (Zhang [2008])

IF $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a standard Gaussian matrix, then with probability at least $1 - \exp(-c_0(n-m))$ ℓ_1-minimization is equivalent to ℓ_0-minimization for all $\mathbf{x}$:

$$\|\mathbf{x}\|_0 < \frac{c_1^2}{4} \frac{m}{1 + \log(n/m)}$$

where $c_0, c_1 > 0$ are constants independent of m and n.

From **Wotao Yin: Sparse Optimization Lecture: Sparse Recovery Guarantees**

How to read guarantees

Some basic aspects that distinguish different types of guarantees:

- Recoverability (exact) vs stability (inexact)

All the above three examples are about exact recoverabiity, without considering noisy or nearly sparse signals.

- General $\mathbf{A}$ or special $\mathbf{A}$?

The third is special A

- Universal (all sparse vectors) or instance (certain sparse vector(s))?
- General optimality? or specific to model / algorithm?
- Required property of $\mathbf{A}$: spark, RIP, coherence, NSP, dual certificate?
- If randomness is involved, what is its role?

The third is involved with randomness

- Condition/bound is tight or not? Absolute or in order of magnitude?

Restricted isometry property (RIP)

Definition (Candes and Tao [2005])

Matrix $\mathbf{A}$ obeys the restricted isometry property (RIP) with constant δ_s if

$$(1-\delta_s)\|\mathbf{c}\|_2^2 \le \|\mathbf{A}\mathbf{c}\|_2^2 \le (1+\delta_s)\|\mathbf{c}\|_2^2$$

for all s-sparse vectors $\mathbf{c}$.

Theorem (Candes and Tao [2006])

If $\mathbf{x}$ is k-sparse and $\mathbf{A}$ satisfies $\delta_{2k}+\delta_{3k}<1$, then $\mathbf{x}$ is the unique ℓ_1 minimizer.

Comments:

- RIP needs a matrix to be properly scaled

  Columns normalized to unit-length, etc.

- the tight RIP constant of a *given matrix* $\mathbf{A}$ is difficult to compute