Coordinate descent

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

December 30, 2023

Outline

Algorithm description

Problem setting

consider the problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{C} := \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_m$$

where x can be partitioned as blocks: $x = (x_1, x_2, \ldots, x_m)$

when f has loose coupling, it is possible to minimize f w.r.t. each block x_k

e.g., while other of x_j's are fixed, minimization w.r.t. x_k becomes fairly easy

example: QP with a box constraint

$$\text{minimize} \quad (1/2)x^T P x + q^T x \quad \text{subject to} \quad l \preceq x \preceq u$$

e.g., appears in dual of soft-margin SVM

Examples of suitable problem structures

1. dual of QP: minimize $(1/2)x^TPx + q^Tx$ subject to $Ax \preceq b$ is

$$\text{minimize} \quad (1/2)\lambda^T G\lambda + s^T\lambda \quad \text{subject to} \quad \lambda \succeq 0$$

where $G = AP^{-1}A^T$ and $s = b + AP^{-1}q$ (dual has simpler constraints)

2. nonnegative matrix factorization: not jointly convex but bi-convex

$$\underset{X,Z}{\text{minimize}} \quad \|ZX - A\|_F^2 \quad \text{subject to} \quad Z \geq 0, X \geq 0$$

factorize A into a product of two matrices having non-negative entries

3. given closed convex sets $\mathcal{C}_i$ for $i = 1, 2, \ldots, m$ and find a point in their intersections – equivalent to the problem with variables $x, y_1, y_2, \ldots, y_m \in \mathbf{R}^n$

$$\text{minimize} \quad (1/2)\sum_{i=1}^m \|y_i - x\|^2 \quad \text{subject to} \quad x \in \mathbf{R}^n, \ \ y_i \in \mathcal{C}_i, \ \ i = 1, 2, \ldots, m$$

notes: in these examples, calculation of minimum along each block can be simplified

Block coordinate descent algorithm (BCD)

denote x_i^+, x_i the next and current iteration of the ith block of x (out of m blocks)

repeats the following m-updates in **cyclic order**

$$
\begin{array}{rcl}
x_1^+ & = & \underset{z \in \mathcal{C}_1}{\operatorname{argmin}} \; f(z, x_2, x_3, \ldots, x_m) \\
x_2^+ & = & \underset{z \in \mathcal{C}_2}{\operatorname{argmin}} \; f(x_1^+, z, x_3, \ldots, x_m) \\
& \vdots & \\
x_i^+ & = & \underset{z \in \mathcal{C}_i}{\operatorname{argmin}} \; f(x_1^+, \ldots, x_{i-1}^+, z, x_{i+1}, \ldots, x_m) \\
& \vdots & \\
x_m^+ & = & \underset{z \in \mathcal{C}_m}{\operatorname{argmin}} \; f(x_1^+, x_2^+, \ldots, x_{m-1}^+, z)
\end{array}
$$

each iteration the cost is minimized w.r.t. each block coordinate

Examples

Example: box-constrained QP

given $q \in \mathbf{R}^n, P \succ 0$ with p_i^T as each row of P

$$\underset{x}{\text{minimize}} \quad (1/2)x^T P x + q^T x \quad \text{subject to} \quad l \preceq x \preceq u$$

minimizing along x_i is simple; first finding the zero-gradient condition w.r.t. x_i

$$\frac{\partial f}{\partial x_i} = (Px)_i + q_i = 0 \quad \Rightarrow \quad p_i^T x + q_i = 0 \Rightarrow \bar{x}_i = -\frac{1}{p_{ii}}\left(q_i + \sum_{k \neq i} p_{ik} x_k\right)$$

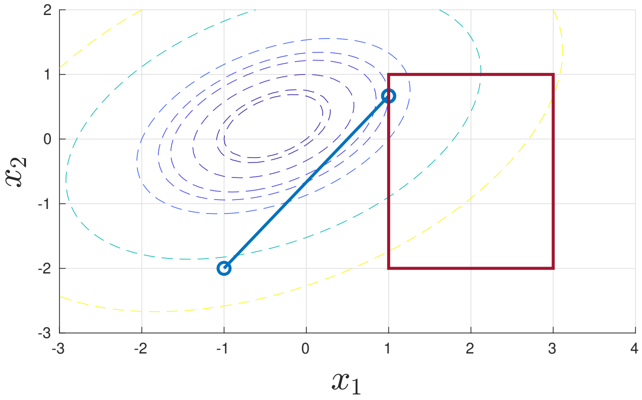
with the box constraint on ith coordinate: $l_i \leq x_i \leq u_i$ then the minimizer is

$$x_i^* = \Pi_{\text{box}}(\bar{x}_i) = \begin{cases} u_i, & \bar{x}_i > u_i \\ \bar{x}_i, & l_i \leq \bar{x}_i \leq u_i \\ l_i, & \bar{x}_i < l_i \end{cases}$$

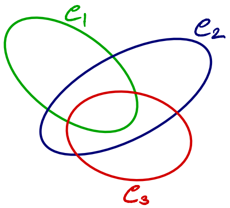
the update on the ith coordinate is simply a projection onto a box

example: results show with $x^{(0)} = (-1, -2)$

$$f(x) = x^T \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} x + \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T x, \quad \begin{bmatrix} 1 \\ -2 \end{bmatrix} \preceq x \preceq \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$



Example: parallel projections



given closed convex sets $\mathcal{C}_i$ for $i = 1, 2, \ldots, m$
find a point in their intersections

$$\begin{array}{ll} \text{minimize} & (1/2)\sum_{i=1}^m \|y_i - x\|_2^2 \\ \text{subject to} & x \in \mathbf{R}^n, \quad y_i \in \mathcal{C}_i, \quad i = 1, 2, \ldots, m \end{array}$$

with variables $y_i, x \in \mathbf{R}^n$

- when x is fixed, the updates on y_i's are separable (cyclic order is then not needed)

$$y_i^+ = \Pi_{\mathcal{C}_i}(x), \quad i = 1, 2, \ldots, m$$

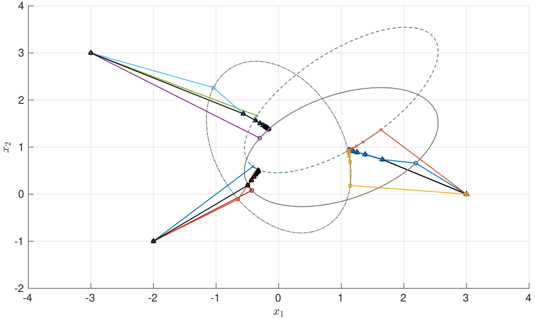
- after y_i's are updated and fixed, the minimization w.r.t. x is just averaging

$$x^+ = \frac{1}{m}\sum_{i=1}^m y_i^+$$

example: $\mathcal{C}_i$ is an ellipsoid of the form: $0.5(x-c_i)^T P_i (x-c_i) \leq \alpha_i$

$$P_1 = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}, P_2 = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}, P_3 = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$$

$$c_1 = (1,1), c_2 = (1,2), c_3 = (0,1), \alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 3$$



projection onto an ellipsoid is not trivial

but 3 projections can be done in parallel

results shown with three different initial points (three lines are y_1, y_2, y_3 sequences)

Convergence

Convergence in convex case

assumptions:

- f is convex and differentiable, $\mathcal{C}_i$'s are closed and convex
- for each $x = (x_1, \dots, x_m) \in \mathcal{C}$ and each i

$$f(x_1, x_2, \dots, x_{i-1}, z, x_{i+1}, \dots, x_m) \quad \text{viewed as a function of } z$$

attains a unique minimum over $\mathcal{C}_i$

results: every limit points of sequence generated by BCD minimizes f over $\mathcal{C}$

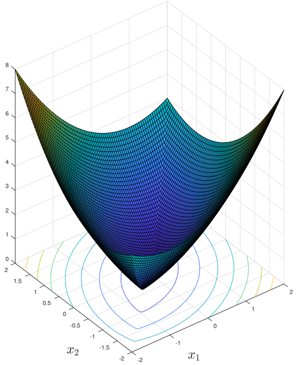
proof follows from D.P. Bertsekas (convex optimization algorithms) on page 371

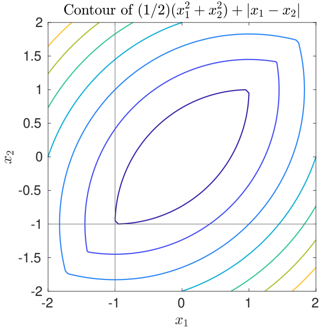
- show that a limit point $\bar{x}$ satisfies $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0, \; \forall x \in \mathcal{C}$
- BCD may fail to converge for non-smooth f even it is convex

BCD may fail to converge

example: minimize $f(x) = (1/2)(x_1^2 + x_2^2) + |x_1 - x_2|$ (non-differentiable)

$$x^{(0)} = (-1,-1), \quad x^{(1)} = (-1,-1), \ldots, \quad \text{while the optimum is at } x^\star = (0,0)$$

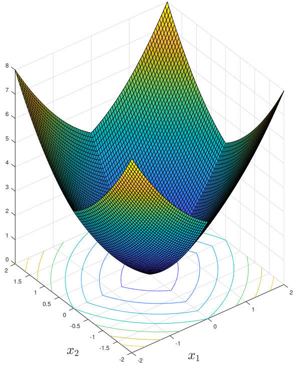


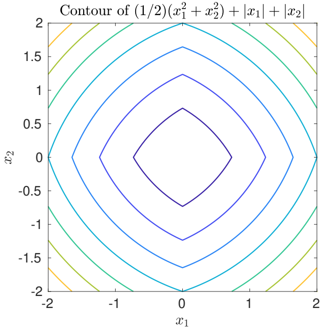


a type of fused lasso where the surface has corners (and sequence is stuck there)

BCD converges for some non-differentiable f

example: minimize $f(x) = (1/2)(x_1^2 + x_2^2) + |x_1| + |x_2|$ (non-differentiable)





the BCD sequences converge to the optimum at $x^\star = (0, 0)$

the non-differentiable part seems to have *some structure* – here it's separable

Convergence for non-smooth convex case

a convergence result from Tseng 2001, Theorem 4.1 (a) – recap in Hastie 2015

assumptions:

- $f(x) = g(x) + \sum_{i=1}^{m} h_i(x_i)$
- g is convex and differentiable, each h_i is convex but can be non-differentiable
- initial level set $S_0 = \{x \mid f(x) \leq f(x^{(0)})\}$ is compact
- f has regularity condition on the directional derivative along Δx

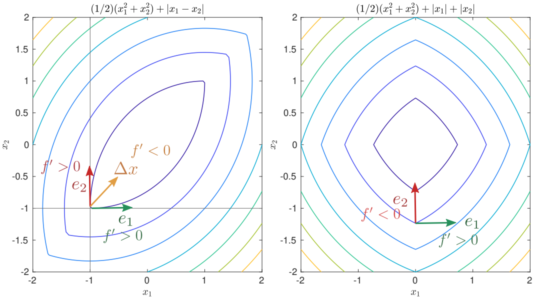
$$f'(x; e_i) \geq 0, \quad i = 1, 2, \ldots, m \quad \Longrightarrow \quad f'(x, \Delta x) \geq 0, \;\; \forall \Delta x \in \mathbf{R}^n$$

f' along each coordinate give sufficient information that moving to other directions will also further increase f

results: every limit point of sequences generated by BCD minimizes f over C

Fused lasso is not regular

BCD only gain information about directions of the form e_j, $j = 1, 2, \ldots, m$



- if reaching a point where f increases along each of *all* e_j's, moving to any other direction should not possibly decrease f – what we called regular
- fused lasso objective is not regular; f increases along both e_1 and e_2 but there are some direction that f decreases
- lasso objective is regular; information where f increases in some direction can be sufficiently obtained from info of f' along *some* e_i

Problems with separable non-smooth parts

$f(x) = g(x) + \sum_{i=1}^m h_i(x_i)$ (the non-differentiable part is separable)

- lasso formulation: minimize $(1/2)\|y - Ax\|_2^2 + \lambda\|x\|_1$
- logistic regression (soft-max cose) with ℓ_q-norm regularization

$$\underset{x}{\text{minimize}} \quad (1/N)\sum_{i=1}^N \log(1 + e^{-y_i z_i^T x}) + \lambda\|x\|_q^q$$

  where $\|x\|_q^q = \sum_{i=1}^m |x_i|^q$, for $0 < q \leq 1$ (non-convex for $q < 1$)
- soft-margin SVM using hinge cost (hinge primal problem)

$$\underset{w}{\text{minimize}} \quad (1/2)\|w\|_2^2 + \lambda\sum_{i=1}^N \max(0, 1 - y_i(x_i^T w + b))$$

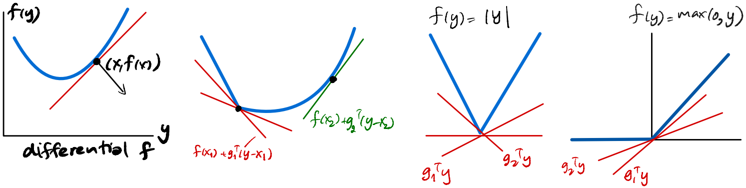
Non-differentiability: subgradient

recall the first-order condition for convexity in f

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall y \in \mathbf{dom}\, f$$

definition: g is a **subgradient** of a convex function f at $x \in \mathbf{dom}\, f$ if

$$f(y) \geq f(x) + g^T (y - x), \quad \forall y \in \mathbf{dom}\, f$$



$\tilde{f}(y) = f(x) + g^T (y - x)$ is an affine function that is a lower bound for $f(y)$ at x

Subdifferential

the concept of subgradients is generalized for for non-differentiable f

- a subgradient of at x is not necessarily unique
- $f(y) = |y|$, subgradient of f at $y = 0$ is any $g \in [-1, 1]$
- $f(y) = \max(0, y)$, subgradient of f at $y = 0$ is any $g \in [0, 1]$
- $f(y) = \|y\|_2$, subgradient of f at $y = 0$ is any g with $\|g\|_2 \leq 1$

  $$f(y) = \|y\|_2 \geq f(0) + g^T(y - 0) = g^T y \text{ when } \|g\|_2 \leq 1$$

  (from Cauchy-Schwarz inequality)
- definition: the **subdifferential** $\partial f(x)$ of f at x is the set of all subgradients

Subgradient calculus

optimality condition for unconstrained problem

$x^\star$ minimizes $f(x)$ if and only if $0 \in \partial f(x^\star)$

$$f(y) \geq f(x^\star) + 0^T(y - x^\star),\ \forall y \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

$f(x^\star)$ is smallest iff 0 is one of the subgradients (follow from the definition of g)

Example: lasso regression

minimize $(1/2)\|y - Ax\|_2^2 + \lambda\|x\|_1$ where $A \in \mathbf{R}^{m\times n}, y \in \mathbf{R}^m, \lambda > 0$ are given

- let a_i, $i = 1, 2, \dots, n$ be columns of A

$$f(x) = (1/2)\|y - (a_1x_1 + a_2x_2 + \cdots + a_nx_n)\|_2^2 + \lambda(|x_1| + |x_2| + \cdots + |x_n|)$$

- minimization of f over x_i (while other x_k's are fixed) is to minimize

$$\tilde{f}(x_i) = (1/2)\|r - a_ix_i\|_2^2 + \lambda|x_i|, \quad r = y - \sum_{k\neq i} a_kx_k \text{ (partial residual)}$$

- optimality condition: zero is one of the subgradients of $\tilde{f}$ w.r.t. to x_i

$$\frac{\partial \tilde{f}}{\partial x_i} = -a_i^Tr + a_i^Ta_ix_i + \lambda s_i = 0, \quad s_i = \begin{cases} 1, & x_i > 0 \\ -1, & x_i < 0 \\ \text{any value in [-1,1]}, & x_i = 0 \end{cases}$$

- three cases at optimality (at kth iteration, and the update of ith block)

$$
\begin{array}{llll}
x_i^\star > 0, & -a_i^T r + \|a_i\|^2 x_i^\star + \lambda \cdot 1 = 0, & \Rightarrow & x_i^\star = \frac{a_i^T r - \lambda}{\|a_i\|^2} \\
x_i^\star < 0, & -a_i^T r + \|a_i\|^2 x_i^\star + \lambda \cdot -1 = 0, & \Rightarrow & x_i^\star = \frac{a_i^T r + \lambda}{\|a_i\|^2} \\
x_i^\star = 0, & -a_i^T r + \|a_i\|^2 \cdot 0 + \lambda \cdot s_i = 0, & \Rightarrow & |a_i^T r| = \lambda |s_i| \leq \lambda
\end{array}
$$

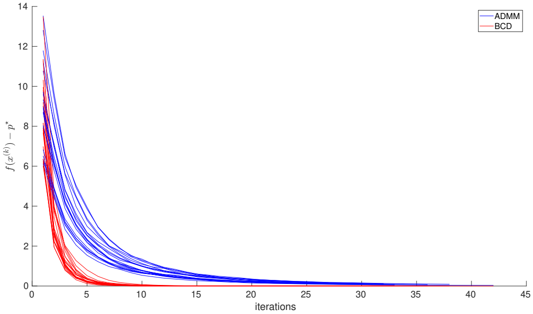
- x_i^+ is then obtained by soft-thresholding operator

$$
x_i^+ = \begin{cases} \frac{a_i^T r - \lambda}{\|a_i\|^2}, & a_i^T r > \lambda \\ \frac{a_i^T r + \lambda}{\|a_i\|^2}, & a_i^T r < -\lambda \\ 0, & |a_i^T r| \leq \lambda \end{cases} \qquad = \frac{1}{\|a_i\|^2} S_\lambda \left(a_i^T (y - \sum_{k \neq i} a_k x_k) \right)
$$

- we apply soft-thresholding to the ith block in cyclic order
- each coordinate update, it takes $\mathcal{O}(m)$ to update r, and $\mathcal{O}(m)$ to update $a_i^T r$; hence, in one cycle, it costs $\mathcal{O}(mn)$ flops

Numerical results of lasso

example: lasso with $A \in \mathbf{R}^{150\times 500}$ and $\lambda = 0.1\lambda_{\max}$ (20 instances)



- all methods were initialized with $x^{(0)} = 0$
- ADMM was implemented with $\rho = 3, \epsilon^{\text{abs}} = 10^{-4}, \epsilon^{\text{rel}} = 10^{-3}$
- BCD stopped when $\|x^+ - x\| \leq 10^{-3}$ (relative difference can be used also)
- both methods had comparable performances in this example

Convergence in non-convex case

assumptions:

- f is continuously differentiable, $\mathcal{C}_i$'s are closed and convex
- for each $x = (x_1, \ldots, x_m) \in \mathcal{C}$ and each i

$$f(x_1, x_2, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_m) \quad \text{viewed as a function of } z$$

  - attains a unique minimum $\bar{z}$ over $\mathcal{C}_i$
  - monotonically non-increasing in the interval from x_i to $\bar{z}$

results: every limit point $\bar{x}$ of sequence generated by BCD satisfies the optimality condition

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0, \quad \forall x \in \mathcal{C}$$

no convexity in f is needed but extra condition on monotonicity is required

Variants

Variants of coordinate descent

more literature and further reading on

- applying the coordinate descent in the context of dual problem (where constraint involves $\mathbf{R}_+^n$
- combination of coordinate descent with the proximal algorithm
- the use of an irregular order instead of a fixed cyclic order (*e.g.*, randomization)

see references in Bertsekas 2015 and Wright 2015

References

1. Chapter 6.5, D.P. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, 2015
2. Chapter 2.7, D.P. Bertsekas, *Nonlinear Programming*, 2nd edition, Athena Scientific, 1999
3. Chapter 12.5, G. Calafiore and L. El Ghaoui, *Optimization Models*, Cambridge University Press, 2014
4. Chapter 5.4, T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity : The Lasso and Generalizations*, CRC Press, 2015
5. P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of optimization theory and applications, 2001
6. S. J. Wright, *Coordinate descent algorithms*, Mathematical Programming, 2015