6th Traditional Summer School for Young Researchers: "Control. Information. Optimization"

Giuseppe Carlo Calafiore

Dipartimento di Automatica e Informatica
Politecnico di Torino – ITALY

Moscow, June 22-29

LECTURE 4

Coordinate minimization methods

Outline

Introduction

- Coordinate descent methods, or more generally block-coordinate descent methods, apply to problems where each variable (or block of variables) is *independently* constrained.
- We consider a special case of a generic minimization problem

$$\min_{x=(x_1,\ldots,x_\nu)} f_0(x) \;:\; x_i \in \mathcal{X}_i, \quad i=1,\ldots,\nu. \tag{1}$$

  In words, the variable x can be decomposed into ν blocks $x_1,\ldots,x_\nu$, and each block x_i is independently constrained to belong to the set $\mathcal{X}_i$.
- Coordinate descent methods are based on iteratively minimizing with respect to one block, with all the other blocks being fixed.
- If $x^{(k)}=(x_1^{(k)},\ldots,x_\nu^{(k)})$ denotes the value of the decision variable at iteration k, partial minimization problems of the form

$$\min_{x_i\in\mathcal{X}_i} f_0(x_1^{(k)},\ldots,x_{i-1}^{(k)},x_i,x_{i+1}^{(k)},\ldots,x_\nu^{(k)}), \tag{2}$$

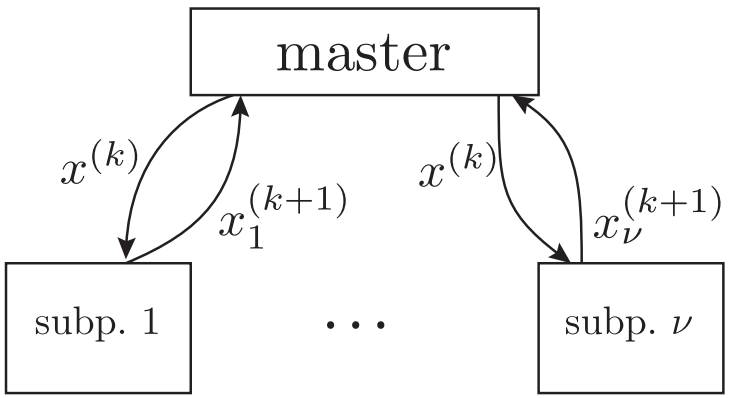
  are solved.
- Different methods ensue, based on how exactly we form the next iterate.

The Jacobi method

- In the Jacobi method, we solve all the partial minimization problems (2), and then update all the blocks *simultaneously*. That is, for every $i = 1, \ldots, \nu$, we set

$$x_i^{(k+1)} = \arg\min_{x_i \in \mathcal{X}_i} f_0(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k)}, \ldots, x_\nu^{(k)}).$$

- The scheme is depicted in the following figure.



- Convergence is not guaranteed, even for convex and smooth objectives.

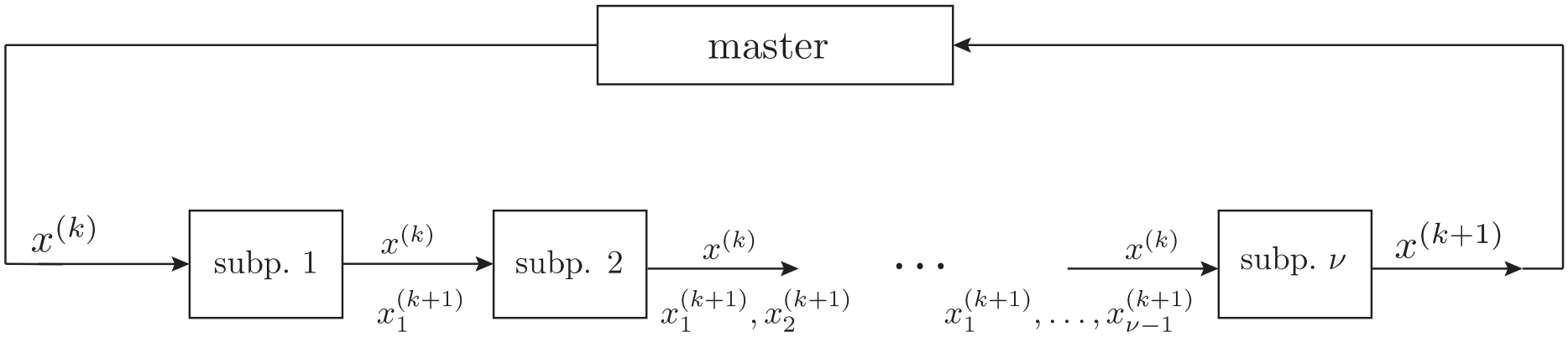
The Gauss–Seidel method

- The standard (block) coordinate minimization method (BCM), also known as the Gauss–Seidel method, works similarly to the Jacobi method, but in this algorithm the variable blocks are updated *sequentially*, according to the recursion

$$x_i^{(k+1)} = \arg\min_{x_i \in \mathcal{X}_i} f_0(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i, x_{i+1}^{(k)}, \ldots, x_\nu^{(k)}), \tag{3}$$

  for $i = 1, \ldots, \nu$.

- The scheme is depicted in the following figure.



The Gauss–Seidel method

Convergence.

Theorem 1

- *Assume f_0 is convex and continuously differentiable on $\mathcal{X}$. Moreover, let f_0 be strictly convex in x_i, when the other variable blocks x_j, $j \neq i$ are held constant.*
- *If the sequence $\{x^{(k)}\}$ generated by the BCM algorithm is well defined, then every limit point of $\{x^{(k)}\}$ converges to an optimal solution of problem (1).*

The Gauss–Seidel method

Convergence.

- The sequential block-coordinate descent method may fail to converge, in general, for non-smooth objectives, even under convexity assumptions.
- An important exception, however, arises when f_0 is a composite function which can be written as the sum of a convex and differentiable function ϕ and a separable convex (but possibly non-smooth) term, that is

$$f_0(x) = \phi(x) + \sum_{i=1}^{\nu} \psi_i(x_i), \tag{4}$$

  where ϕ is convex and differentiable, and ψ_i, $i = 1, \ldots, \nu$ are convex.
- Notice that this setup includes the possibility of convex independent constraints on the variables of the form $x_i \in \mathcal{X}_i$, since functions ψ_i may include a term given by the indicator function of the set $\mathcal{X}_i$.
- The structure (4) also includes various ℓ_1-norm regularized problems, such as the LASSO, for which convergence of sequential coordinate descent methods is thus guaranteed.

The Gauss–Seidel method

Convergence.

Theorem 2

- *Let $x^{(0)} \in \mathcal{X}$ be an initial point for the BCM, and assume that the level set $S_0 = \{x : f_0(x) \leq f_0(x^{(0)})\} \in \operatorname{int} \operatorname{dom} f_0$ is compact.*
- *Assume further that f_0 has the form (4), where ϕ is convex and differentiable on S_0, and ψ_i, $i = 1, \ldots, \nu$, are convex.*
- *Then, every limit point of the sequence $\{x^{(k)}\}$ generated by the BCM converges to an optimal solution of problem (1).*

We next sketch the proof of this result.

Minima and coordinate-wise minima

- Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex function taking values $f(x) = f(x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}^{n_i}$ is a block of variables of dimension $n_i \geq 1$, with $\sum_{i=1}^{n} n_i = N$. We assume that $\mathrm{dom}\, f$ is open, and we denote with $E_i \in \mathbb{R}^{N,n_i}$ a stacked block matrix where the j-th block is a zero block of dimension $n_j \times n_i$, for $j \neq i$, and it is an identity block of dimension $n_i \times n_i$, for $j = i$.
- A point $z \in \mathrm{dom}\, f$ is a *coordinate-wise minimum point* of f, if

$$f(z + E_i \xi_i) \geq f(z), \quad \forall \xi_i \in \mathbb{R}^{n_i},\ i = 1, \ldots, n.$$

- $z \in \mathrm{dom}\, f$ is a *minimum point* of f, if

$$f(z + w) \geq f(z), \quad \forall w \in \mathbb{R}^N.$$

Minima and coordinate-wise minima

- For convex f, it is well known that $z \in \operatorname{dom} f$ is a minimum point of f if and only if

$$f'(z, v) \geq 0, \quad \forall v,$$

where $f'(z, v)$ is the directional derivative of f at z along direction v (we recall that the directional derivative of convex f exists at each $x \in \operatorname{int} \operatorname{dom} f$, even if f does not admit a standard gradient):

$$f'(z, v) = \lim_{\lambda \to 0_+} \frac{f(z + \lambda v) - f(z)}{\lambda} = \max_{g \in \partial f(x)} v^\top g,$$

where $\partial f(x)$ is the subdifferential of f at x.

- Also, $z \in \operatorname{dom} f$ is a coordinate-wise minimum point of f if and only if

$$f'(z, E_i v_i) \geq 0, \quad \forall v_i, \ i = 1, \ldots, n.$$

- Clearly, if z is a minimum point of f, it is also a coordinate-wise minimum point. The converse is not true, in general. A relevant exception is discussed next.

Differentiable plus separable structure

Assumption 1

We assume that f has the following form:

$$f(x) = f_0(x) + \sum_{i=1}^{n} \varphi_i(x_i), \tag{5}$$

where f_0 is convex and differentiable on $\mathrm{dom}\, f$ (which is assumed to be open), and φ_i are convex (but possibly non differentiable) functions. Moreover, we assume that f is bounded below, and attains its minimum value f^*.

Differentiable plus separable structure

Proposition 1

Let f satisfy Assumption 1, and suppose z is a coordinate-wise minimum point for f. Then, z is a minimum point of f.

Proof. If z is a coordinate-wise minimum point for f, then, for all $v_i \in \mathbb{R}^{n_i}$ and all $i = 1, \ldots, n$, it holds that $f'(z, E_i v_i) \geq 0$. We thus have that

$$\begin{aligned} f'(z, v) &= \nabla f_0(z)^\top v + \sum_i \varphi_i'(z_i, v_i) \\ &= \sum_i \nabla_i f_0(z)^\top v_i + \varphi_i'(z_i, v_i) \\ &= \sum_i f'(z, E_i v_i) \geq 0, \end{aligned} \tag{6}$$

which permits to conclude that z is a minimum point of f (in the above, $\nabla_i f_0(z)$ denotes the block of the gradient of f relative to i-th variable block).

Convergence of BCM method

- If we prove that BCM converges to a coordinate-wise minimum point, then using Proposition 1 we also prove convergence to a minimum point.
- Let $f^{(k)} \doteq f(x(k))$, $k = 0, 1, \ldots$, for some initial point $x(0) \in \mathrm{dom}\, f$. Here, k denotes the "outer" iteration count, after a full serial sweep over all the blocks is performed.
- Since at each "inner" iteration over a the i-th block the function value cannot increase (f is minimized wrt the i-th block), we have that $f^{(0)} < \infty$, and $f^{(k+1)} \leq f^{(k)}$.
- The sequence $\{f^{(k)}\}$ is thus non-increasing. Further, since f is bounded below, the sequence converges to some limit $\bar{f}$, that is: $\lim_{k\to\infty} f^{(k)} = \bar{f}$, hence $\lim_{k\to\infty} f^{(k+1)} - f^{(k)} = 0$.
- This means that the algorithm reaches asymptotically a point for which the objective connot be improved by element-wise minimization, i.e., it reaches a coordinate-wise minimum point.
- By Proposition 1 we conclude that the BCM algorithm converges to a minimum point of f. □

Coordinate minimization for the LASSO

Exercise.

Code in Matlab a coordinate minimization algorithm for solving the LASSO problem

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2}\|Ax - y\|_2^2 + \lambda \|x\|_1,$$

where $A \in \mathbb{R}^{m,n}$ is a given matrix, $y \in \mathbb{R}^m$ is a given vector, and $\lambda \geq 0$ is an assigned scalar tradeoff parameter.

Hints...

- If x is the current point, then the i-th coordinate minimization problem takes the form (all but the i-variable are fixed to the values in x, and we minimize with respect to x_i)

$$\min_{x_i \in \mathbb{R}} \ \frac{1}{2}\|a_i x_i - y^{(i)}\|_2^2 + \lambda |x_i|_1 + \lambda \|x_{(-i)}\|_1,$$

  where a_i is the i-th column of A, $x_{(-i)}$ is a vector obtained from x by fixing the i-th entry to zero, and $y^{(i)} \doteq y - Ax_{(-i)}$.

- Prove that the above uni-variate minimization problem has the following optimal solution:

$$x_i^* = \begin{cases} 0 & \text{if } |a_i^\top y^{(i)}| \leq \lambda, \\ \xi_i - \operatorname{sgn}(\xi_i)\frac{\lambda}{\|a_i\|_2^2} & \text{if } |a_i^\top y^{(i)}| > \lambda, \end{cases}$$

  where

$$\xi_i \doteq \frac{a_i^\top y^{(i)}}{\|a_i\|_2^2}$$

  corresponds to the solution of the problem for $\lambda = 0$. Verify that this solution can be expressed more compactly as $x_i^* = \operatorname{sthr}_{\lambda/\|a_i\|_2^2}(\xi_i)$, where sthr is the *soft threshold* function.