Optimization in Machine Learning: Coordinate Descent/Minimization Method

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Spring 2023)
http://cs182.sist.shanghaitech.edu.cn

Block Coordinate Descent

- Consider the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X} \tag{1}$$

  where each $\mathcal{X}$ is closed, non-empty, and convex.

- Suppose the variable $\mathbf{x}$ can be decomposed into m blocks, i.e., $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$ with $\mathbf{x}_i \in \mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ for $i = 1, \ldots, m$ where $\mathcal{X}_i$ is closed, non-empty, and convex, and $\sum_i n_i = n$, the above problem becomes

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \quad \text{s.t. } \mathbf{x}_i \in \mathcal{X}_i, \forall i \tag{2}$$

Block Coordinate Descent

- **BCD Algorithm:**
  1: Find a feasible point $\mathbf{x}^0 \in \mathcal{X}$ and set $r = 0$
  2: **repeat**
  3: $\quad r = r + 1, i = (r - 1 \bmod m) + 1$
  4: $\quad$ Let $\mathbf{x}_i^\star \in \arg\min_{\mathbf{x} \in \mathcal{X}_i} f\left(\mathbf{x}_1^{r-1}, \dots, \mathbf{x}_{i-1}^{r-1}, \mathbf{x}, \mathbf{x}_{i+1}^{r-1}, \dots, \mathbf{x}_m^{r-1}\right)$
  5: $\quad$ Set $\mathbf{x}_i^r = \mathbf{x}_i^\star$ and $\mathbf{x}_k^r = \mathbf{x}_k^{r-1}, \forall k \neq i$
  6: **until** some convergence criterion is met
- Merits of BCD
  1. Each subproblem is much easier to solve, or even has a closed-form solution;
  2. The objective value is non-increasing along the BCD updates;
  3. It allows parallel or distributed implementations.

Applications — $\ell_2 - \ell_1$ Optimization Problem

- Let us revisit the $\ell_2 - \ell_1$ problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{y} - \mathbf{Ax}\|_2^2 + \mu\|\mathbf{x}\|_1 . \tag{3}$$

- Apart from MM, BCD is another efficient approach to solve (3):
  - Optimize x_k while fixing $x_j = x_j^r, \forall j \neq k$:

$$\min_{x_k} f_k(x_k) \triangleq \frac{1}{2}\|\underbrace{\mathbf{y} - \sum_{j\neq k}\mathbf{a}_j x_j^r}_{\triangleq\overline{\mathbf{y}}} - \mathbf{a}_k x_k\|_2^2 + \mu|x_k| .$$

  - The optimal x_k has a closed form:

$$x_k^\star = \mathrm{soft}\left(\mathbf{a}_k^\top\overline{\mathbf{y}}/\|\mathbf{a}_k\|^2, \mu/\|\mathbf{a}_k\|^2\right),$$

    where $\mathrm{soft}(u,a) \triangleq \mathrm{sign}(u)\max\{|u|-a, 0\}$ denotes a *soft-thresholding* operation.
  - Cyclically update $x_k, k = 1, \ldots, n$ until convergence.

Applications — Low-Rank Matrix Completion

- In the matrix factorization lecture, we have introduced the low-rank matrix completion problem, which has huge potential in sales recommendation.
- For example, we would like to predict how much someone is going to like a movie based on its movie preferences:

$$\begin{array}{ccc} & \text{movies} & \\ \mathbf{M} = & \begin{bmatrix} 2 & 3 & 1 & ? & ? & 5 & 5 \\ 1 & ? & 4 & 2 & ? & ? & ? \\ ? & 3 & 1 & ? & 2 & 2 & 2 \\ ? & ? & ? & 3 & ? & 1 & 5 \\ 2 & ? & 4 & ? & ? & 5 & 3 \end{bmatrix} & \text{users} \end{array}$$

- $\mathbf{M}$ is assumed to be of low rank, as only a few factors affect users' preferences.

$$\min_{\mathbf{W}\in\mathbb{R}^{m\times n}} \cancel{\mathrm{rank}(\mathbf{W})} \; \|\mathbf{W}\|_* \quad \text{s.t. } w_{ij} = m_{ij}, \forall (i,j) \in \mathbf{\Omega}.$$

- An alternative low-rank matrix completion formulation [3]:

$$(\triangle)\ \min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \frac{1}{2}\|\mathbf{XY}-\mathbf{Z}\|_{\mathsf{F}}^2 \quad \text{s.t. } z_{ij}=m_{ij}, \forall(i,j)\in\mathbf{\Omega},$$

  where $\mathbf{X}\in\mathbb{R}^{M\times L}, \mathbf{Y}\in\mathbb{R}^{L\times N}, \mathbf{Z}\in\mathbb{R}^{M\times N}$, and L is an estimate of min. rank.
- Advantage of adopting $(\triangle)$: When BCD is applied, each subproblem of $(\triangle)$ has a closed-form solution:

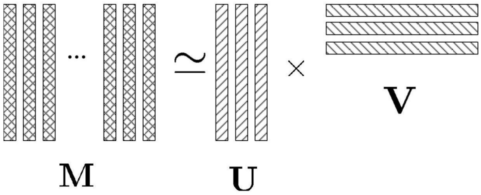
$$\mathbf{X}^{r+1} = \mathbf{Z}^r\mathbf{Y}^{r\top}\left(\mathbf{Y}^r\mathbf{Y}^{r\top}\right)^{\dagger},$$
$$\mathbf{Y}^{r+1} = \left(\mathbf{X}^{r+1\top}\mathbf{X}^{r+1}\right)^{\dagger}\left(\mathbf{X}^{r+1\top}\mathbf{Z}^r\right),$$
$$\left[\mathbf{Z}^{r+1}\right]_{i,j} = \begin{cases} \left[\mathbf{X}^{r+1}\mathbf{Y}^{r+1}\right]_{i,j}, & \text{for } (i,j)\notin\mathbf{\Omega} \\ m_{i,j}. & \text{for } (i,j)\in\mathbf{\Omega} \end{cases}$$

Applications — Non-negative Matrix Factorization (NMF)

- NMF is concerned with the following problem [2]:

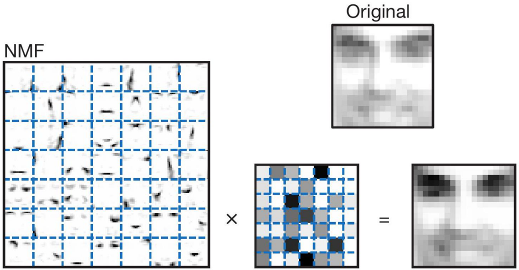
$$\min_{\mathbf{U}\in\mathbb{R}^{m\times k},\mathbf{V}\in\mathbb{R}^{k\times n}} \|\mathbf{M}-\mathbf{UV}\|_F^2 \quad \text{s.t. } \mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}, \tag{4}$$

  where $\mathbf{M}\geq\mathbf{0}$.
- Usually $k \ll \min(m,n)$ or $mk+nk \ll mn$, so NMF can be seen as a linear dimensionality reduction technique for non-negative data.



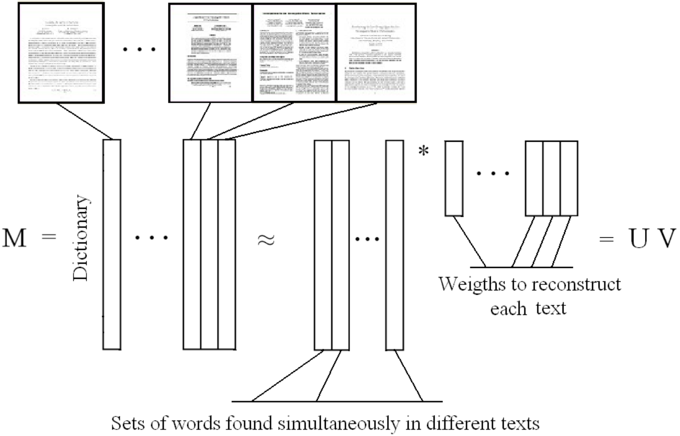
NMF Examples

- Image processing:
  - $\mathbf{U} \geq \mathbf{0}$ constraints the basis elements to be non-negative.
  - $\mathbf{V} \geq \mathbf{0}$ imposes an additive reconstruction.



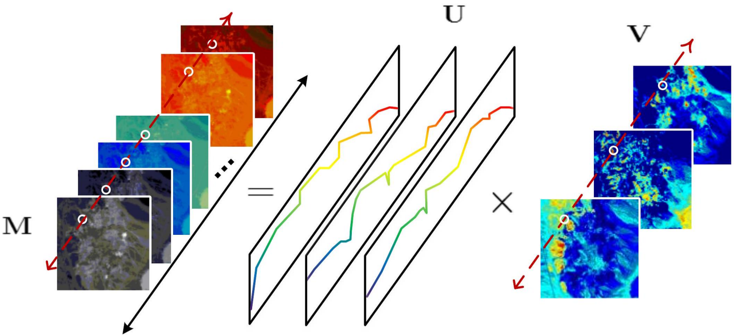
The basis elements extract facial features such as eyes, noses, and lips.

- Text mining:



- Basis elements allow to recover different topics;
- Weights allow to assign each text to its corresponding topics.

- Hyperspectral unmixing:



  - Basis elements **U** represent different materials;
  - Weights **V** allow to know which pixel contains which material.

- Let's turn back to the NMF problem:

$$\min_{\mathbf{U}\in\mathbb{R}^{m\times k},\mathbf{V}\in\mathbb{R}^{k\times n}} \|\mathbf{M}-\mathbf{U}\mathbf{V}\|_F^2 \quad \text{s.t. } \mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}. \tag{5}$$

- Without "$\geq \mathbf{0}$" constraints, the optimal $\mathbf{U}^\star$ and $\mathbf{V}^\star$ can be obtained by SVD.
- With "$\geq \mathbf{0}$" constraints, problem (5) is generally NP-hard.
- When fixing $\mathbf{U}$ (resp. $\mathbf{V}$), problem (5) is convex w.r.t. $\mathbf{V}$ (resp. $\mathbf{U}$).
- For example, for a given $\mathbf{U}$, the i th column of $\mathbf{V}$ is updated by solving the following NNLS problem:

$$\min_{\mathbf{V}(:,i)\in\mathbb{R}^k} \|\mathbf{M}(:,i)-\mathbf{U}\mathbf{V}(:,i)\|_2^2, \quad \text{s.t. } \mathbf{V}(:,i)\geq\mathbf{0}. \tag{6}$$

BCD Algorithm for NMF:

Initialize $\mathbf{U} = \mathbf{U}^0, \mathbf{V} = \mathbf{V}^0$ and $r = 0$;
repeat
    solve the NNLS problem

$$\mathbf{V}^\star \in \arg \min_{\mathbf{V} \in \mathbb{R}^{k \times n}} \|\mathbf{M} - \mathbf{U}^r \mathbf{V}\|_F^2, \quad \text{s.t. } \mathbf{V} \geq \mathbf{0};$$

    $\mathbf{V}^{r+1} = \mathbf{V}^\star$;
    solve the NNLS problem

$$\mathbf{U}^\star \in \arg \min_{\mathbf{U} \in \mathbb{R}^{m \times k}} \left\|\mathbf{M} - \mathbf{U}\mathbf{V}^{r+1}\right\|_F^2, \quad \text{s.t. } \mathbf{U} \geq \mathbf{0};$$

    $\mathbf{U}^{r+1} = \mathbf{U}^\star$;
    $r = r + 1$;
until some convergence criterion is met.

BCD Convergence

- The idea of BCD is to divide and conquer. However, there is no free lunch; BCD may get stuck or converge to some point of no interest.

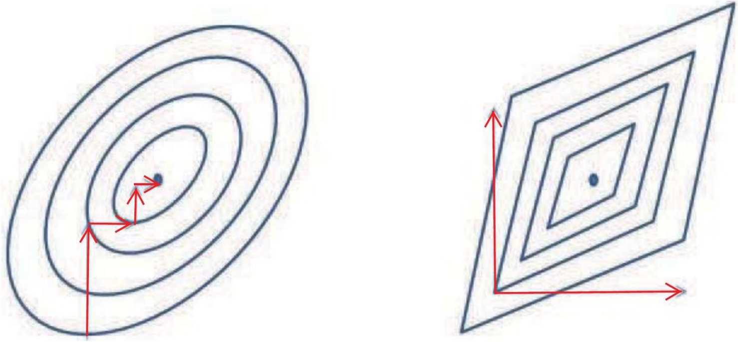


Figure: BCD for smooth/non-smooth minimization.

BCD Convergence

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m \subseteq \mathbb{R}^n \tag{7}$$

- A well-known BCD convergence result is due to Bertsekas:

Theorem

Suppose that f is continuously differentiable over the convex closed set $\mathcal{X}$. Furthermore, suppose that for each i

$$g_i(\xi) \triangleq f\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}, \xi, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_m\right),$$

is strictly convex. Let $\{\mathbf{x}^r\}$ be the sequence generated by BCD method. Then every limit point of $\{\mathbf{x}^r\}$ is a stationary point of problem (7).

- If $\mathcal{X}$ is (convex) compact, i.e., closed and bounded, then strict convexity of $g_i(\xi)$ can be relaxed to having a unique optimal solution.

Generalization of Bertsekas' Convergence Result

- Generalization 1: Relax Strict Convexity to Strict Quasiconvexity [1][1]

Theorem

Suppose that the function f is continuously differentiable and strictly quasiconvex with respect to $\mathbf{x}_i$ on $\mathcal{X}$, for each $i = 1, \ldots, m-2$ and that the sequence $\{\mathbf{x}^r\}$ generated by the BCD method has limit points. Then, every limit point is a stationary point of (7).

- Application: Low-Rank Matrix Completion

$$(\triangle) \ \min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \frac{1}{2}\|\mathbf{X}\mathbf{Y}-\mathbf{Z}\|_{\mathrm{F}}^2 \quad \text{s.t. } z_{ij} = m_{ij}, \forall (i,j) \in \mathbf{\Omega}.$$

  - $m = 3$ and $(\triangle)$ is strictly convex w.r.t. $\mathbf{Z} \Longrightarrow$ BCD converges to a stationary point.

[1] f is strictly quasiconvex w.r.t. $\mathbf{x}_i \in \mathcal{X}_i$ on $\mathcal{X}$ if for every $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y}_i \in \mathcal{X}_i$ with $\mathbf{y}_i \neq \mathbf{x}_i$ we have

$$f(\mathbf{x}_1, \ldots, t\mathbf{x}_i + (1-t)\mathbf{y}_i, \ldots, \mathbf{x}_m) < \max\{f(\mathbf{x}), f(\mathbf{x}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{x}_m)\}, \forall t \in (0,1).$$

- Generalization 2: Without Solution Uniqueness

Theorem

Suppose that f is continuously differentiable, and that $\mathcal{X}$ is convex and closed. Moreover, if there are only two blocks, i.e., $m = 2$, then every limit point generated by BCD is a stationary point of f.

- Application: NMF

$$\min_{\mathbf{U}\in\mathbb{R}^{m\times k},\mathbf{V}\in\mathbb{R}^{k\times n}} \|\mathbf{M}-\mathbf{U}\mathbf{V}\|_{\mathsf{F}}^2 \quad \text{s.t. } \mathbf{U}\geq\mathbf{0},\mathbf{V}\geq\mathbf{0}.$$

- Alternating NNLS converges to a stationary point of the NMF problem, since
  - the objective is continuously differentiable;
  - the feasible set is convex and closed;
  - $m = 2$.

L. Grippo and M. Sciandrone.
On the convergence of the block nonlinear gauss–seidel method under convex constraints.
Operations Research Letters, 26(3):127–136, 2000.

Daniel Lee and H. Sebastian Seung.
Algorithms for non-negative matrix factorization.
In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Zaiwen Wen, Wotao Yin, and Yin Zhang.
Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm.
Mathematical Programming Computation, 4(4):333–361, 2012.