**Optimization Theory and Algorithm II** **October 20, 2022**

Lecture 12

*Lecturer:Xiangyu Chang* *Scribe: Xiangyu Chang*

*Edited by: Xiangyu Chang*

# 1 Block Coordinate Descent

## 1.1 Motivation

Let us recall the SGD and FedAvg. They all consider a big data setting. For example, the ERM problem (finite-summation optimization):

$$\min_{\mathbf{x}} \ \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x};\mathbf{z}_i).$$

Using GD,

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{s_t}{m}\sum_{i=1}^{m}\nabla f(\mathbf{x}^t,\mathbf{z}_i).$$

The summation is huge due to the big data setting. The basic idea is the "sample decomposition". SGD and FedAvg are the two typical examples.

Another type problem is sightly different, which involves many decision variables called "high-dimensional problems".

**Example 1** *For the least squares problem,*

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2,$$

*where* $\mathbf{x}^* = (A^\top A)^{-1}A^\top\mathbf{b}$. *Computing* $(A^\top A)^{-1}$, *we need* $O(n^3)$ *which is determined by the number of decision. variables. If* $n >> m$, *then* $A^\top A$ *is not invertible, the we have to use GD,*

$$\mathbf{x}^{t+1} = (I - s_t A^\top A)\mathbf{x}^t + s_t A^\top\mathbf{b}.$$

*However, computing* $A^\top A$ *needs* $O(n^2m)$ *computations and* $O(n^2)$ *storage which is prohibited for large* $n$.

**Example 2** *Let us consider LASSO problem again.*

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1 = \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 + \sum_{j=1}^{n}|x_j|.$$

*In this model, maybe* $n$ *is so large.*

**General Formulation:**

$$\min_{\mathbf{x}} f(\mathbf{x}) = f(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_K) + \sum_{k=1}^{K} r_k(\mathbf{x}_k), \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^{n_k}$ and $\sum_k n_k = n$ which means the decision variables are decomposed into $K$ groups. If $K = n$, then $\mathbf{x}_k \in \mathbb{R}$. In this part, we assume that $f \in C^1$ and $r_k, k \in [K]$ are convex.

### 1.1.1 BCD

Suppose that we have an iterative algorithm to obtain $\mathbf{x}^t$, then define that

$$f_k^t(\mathbf{x}_k) := f(\mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1}, \ldots, \mathbf{x}_{k-1}^{t+1}, \mathbf{x}_k, \mathbf{x}_{k+1}^t, \ldots, \mathbf{x}_K^t) = f(\mathbf{x}_{<k}^{t+1}, \mathbf{x}_k, \mathbf{x}_{>k}^t). \tag{2}$$

Solve a sub-optimization problem of Eq.(1).

(i) $\mathbf{x}_k^{t+1} = \arg\min_{\mathbf{x}_k}\{f_k^t(\mathbf{x}_k) + r_k(\mathbf{x}_k)\}$.

(ii) $\mathbf{x}_k^{t+1} = \arg\min_{\mathbf{x}_k}\{f_k^t(\mathbf{x}_k) + \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^t\|^2 + r_k(\mathbf{x}_k)\}$.

(iii) $\mathbf{x}_k^{t+1} = \arg\min_{\mathbf{x}_k}\{\frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^t\|^2 + \langle\nabla f_k^t(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_k^t\rangle + r_k(\mathbf{x}_k)\}$.

Using item $\frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^t\|^2$ is to control the $\mathbf{x}^{t+1}$ is not far away from $\mathbf{x}_k^t$ in a certain sense.

**Example 3** *Let us consider the problem*

$$\min f(x, y) = x^2 - 2xy + 10y^2 - 4x - 20y.$$

*If we fix $y$, then $\nabla_x f(x,y) = 2x - 2y - 4 = 0$, that is $x$=$y$+2. If we fix $x$, then $\nabla_y f(x,y) = 20y - 2x - 20 = 0$, that is $y = x/10 + 1$.*

$$\begin{cases} x^{t+1} = y^t + 2, \\ y^{t+1} = x^{(t+1)}/10 + 1. \end{cases}$$

**Algorithm 1** Block Coordinate Descent

1: **Input:** Given a initial starting point $\mathbf{x}^0 = (\mathbf{x}_1^0, \ldots, \mathbf{x}_K^0) \in \mathbb{R}^n$, and $t = 0$
2: **for** $t = 0, 1, \ldots, T$ **do**
3:  **for** $k = 1, \ldots, K$ **do**
4:  Do (i) or (ii) or (iii) for Eq.(1).
5:  **end for**
6: **end for**
7: **Output:** $\mathbf{x}^T$.

**Remark 1**
- *This algorithm is called "Block Coordinate Descent". If $K = n$, it also called "Coordinate Descent".*
- *This algorithm does not always convert to the optimal solution (see Page 393 of textbook).*
- *The related convergence theory can be found in two review papers [**?**, **?**].*

**Example 4** *(Group LASSO)*

*Suppose that $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n = (\mathbf{z}_1, \ldots, \mathbf{z}_K)^\top$ and $\mathbf{z}_k \in \mathbb{R}^{n_k}, \sum_{k=1}^K n_k = n, A = [A_1, A_2, \ldots, A_K] \in \mathbb{R}^{m\times n}$. Then Group LASSO is*

$$\min_{\mathbf{x}} \ \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\sum_{k=1}^K \|\mathbf{z}_k\|_2,$$

*where $\|\mathbf{z}_k\|_2 = \sqrt{\sum_{l=1}^{n_k} z_{kl}^2}$. This is equivalent to*

$$\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{b} - \sum_{k=1}^K A_k\mathbf{z}_k\|^2 + \lambda\sum_{k=1}^K \|\mathbf{z}_k\|_2. \tag{3}$$

*BCD algorithm: Given* $\mathbf{z}_2^t, \ldots, \mathbf{z}_K^t$, *then let* $\mathbf{b}^t = \mathbf{b} - \sum_{k=2}^K A_k \mathbf{z}_k^t$. *Then Eq.(3) is equivalent to*

$$\min_{\mathbf{z}_1} \frac{1}{2}\|\mathbf{b}^t - A_1\mathbf{z}_1\|^2 + \lambda\|\mathbf{z}_1\|_2.$$

*If* $\mathbf{z}_1 \neq 0$, *then* $-A_1^\top(\mathbf{b}^t - A_1\mathbf{z}_1) + \lambda\frac{\mathbf{z}_1}{\|\mathbf{z}_1\|_2} = 0$, *so,*

$$\mathbf{z}_1 = (A_1^\top A_1 + \frac{\lambda I}{\|\mathbf{z}_1\|_2})^{-1} A_1^\top \mathbf{b}^t.$$

*The iterative step is*

$$\mathbf{z}_1^{t+1} \leftarrow (A_1^\top A_1 + \frac{\lambda I}{\|\mathbf{z}_1^t\|_2})^{-1} A_1^\top \mathbf{b}^t.$$

*If* $\mathbf{z}_1 = 0$, *then* $0 \in \partial(\frac{1}{2}\|\mathbf{b}^t - A_1\mathbf{z}_1\|^2 + \lambda\|\mathbf{z}_1\|_2) = -A_1^\top\mathbf{b}^t + \lambda s$, *where* $s \in \partial\|0\|_2 = \{s|\|s\|_2 \leq 1\}$.
*Thus,* $\|A_1^\top\mathbf{b}^t\| \leq \lambda$. *Final update is*

$$\mathbf{z}_1^{t+1} \leftarrow \begin{cases} 0, & \text{if } \|A_1^\top\mathbf{b}^t\| \leq \lambda, \\ (A_1^\top A_1 + \frac{\lambda I}{\|\mathbf{z}_1^t\|_2})^{-1} A_1^\top \mathbf{b}^t, & \text{otherwise.} \end{cases}$$

**Example 5** *(K-means)*

*Suppose we have a data matrix* $A_{m\times n} = (\mathbf{a}_1^\top, \ldots, \mathbf{a}_m^\top)^\top$. *We introduce a corresponding binary indicator variable* $r_{ik} \in \{0, 1\}, i \in [m], k \in [K]$ *to describe which of the* $k$ *clusters the data point* $\mathbf{a}_i$ *is assigned. If* $\mathbf{a}_i$ *is assigned to cluster* $k$, *then* $r_{ik} = 1$, *otherwise* $r_{ik'} = 0, k' \neq k$. *Let* $\mu_k$ *be the mean vector of cluster* $k$, *then the objective function of* $K$*-means is*

$$\min_{\mu_k, r_{ik}} \sum_{i=1}^m \sum_{k=1}^K r_{ik}\|\mathbf{a}_i - \mu_k\|^2 = \ell(R, \mu), \tag{4}$$

*where* $R \in \mathbb{R}^{m\times K}$ *includes all the indicator variables and* $\mu \in \mathbb{R}^{K\times n}$ *includes all* $\mu_k$.

*K-means Algorithm:*

- *Fix* $r_{ik}$, $\nabla_{\mu_k}\ell(R, \mu) = -2\sum_{i=1}^m r_{ik}(\mathbf{a}_i - \mu_k) = 0$, *that is*

$$\mu_k = \frac{\sum_{i=1}^m r_{ik}\mathbf{a}_i}{\sum_{i=1}^m r_{ik}}.$$

- *Fix* $\mu_k$ *then,*

$$r_{ik^*} = \begin{cases} 1, & \text{if } k^* = \arg\min_{1\leq k\leq K} \|\mathbf{a}_i - \mu_k\|^2, \\ 0, & \text{otherwise.} \end{cases}$$

*We further denote* $\mu = (\mu_1^\top, \mu_2^\top, \ldots, \mu_K^\top)^\top \in \mathbb{R}^{K\times n}$ *and* $R = (r_1^\top, \ldots, r_m^\top)^\top \in \mathbb{R}^{m\times K}$, *then the objective function of K-means can be reformulated as:*

$$\min_{R,\mu} \|A - R\mu\|_F^2.$$

*The K-means algorithm first fixes* $R$ *to solve* $\mu$, *then fixes* $\mu$ *to solve* $R$ *respectively.*

Furthermore, K-means can be considered as a "matrix decomposition" problem. Actually, we can find many different matrix decomposition problems can be solved by the BCD algorithm.

**Example 6** *Suppose we know $M$, then the following problem called "non-negative matrix decomposition" [1]:*

$$\min_{X,Y} \frac{1}{2}\|XY - M\|_F^2, \tag{5}$$

$$s.t.\ X \succeq 0, Y \succeq 0. \tag{6}$$

*Let $f(X,Y) = \frac{1}{2}\|XY - M\|_F^2$, then*

$$\frac{\partial f}{\partial X} = (XY - M)Y^\top, \frac{\partial f}{\partial Y} = X^\top(XY - M).$$

*Then the BCD algorithm is*

$$X^{t+1} = \max\{X^t - s_t^X(X^tY^t - M)(Y^t)^\top, 0\}, \tag{7}$$

$$Y^{t+1} = \max\{Y^t - s_t^Y(X^{t+1})^\top(X^{t+1}Y^t - M), 0\}. \tag{8}$$

# References


[1] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.