

# ISyE8813

## Stochastic Subgradient Methods

Renato Monteiro

February 2, 2024

### 1 The problem and assumptions

Let  $\Xi$  denote the support of random vector  $\xi$  and consider the CO problem

$$\phi_* = \min\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\} \quad (1)$$

where:

(A1)  $f, h \in \overline{\text{Conv}}(\mathbb{R}^n)$  are such that  $\text{dom } f \supset \text{dom } h$ ;

(A2) for almost every  $\xi \in \Xi$ , a functional oracle  $F(\cdot, \xi) : \text{dom } h \rightarrow \mathbb{R}$  and a stochastic subgradient oracle  $s(\cdot, \xi) : \text{dom } h \rightarrow \mathbb{R}^n$  satisfying

$$f(x) = \mathbb{E}[F(x, \xi)], \quad f'(x) := \mathbb{E}[s(x, \xi)] \in \partial f(x)$$

for every  $x \in \text{dom } h$  are available;

(A3) there exist constants  $M, L, \sigma \geq 0$  such that

$$\|f'(\tilde{x}) - f'(x)\| \leq 2M + \frac{L}{2}\|\tilde{x} - x\| \quad \forall x, \tilde{x} \in \text{dom } h$$

and

$$\mathbb{E} \left[ \|s(x; \xi) - f'(x)\|^2 \right] \leq \sigma^2 \quad \forall x \in \text{dom } h;$$

(A4) the set of optimal solutions  $X_*$  of (1) is nonempty.

#### Remarks:

1) condition (A2) does not require  $F(\cdot, \xi)$  to be convex.

2) condition (A3) implies that

$$0 \leq f(\tilde{x}) - \ell_f(\tilde{x}; x) \leq 2M\|\tilde{x} - x\| + \frac{L}{2}\|\tilde{x} - x\|^2 \quad (2)$$

where

$$\ell_f(\tilde{x}; x) = f(x) + \langle f'(x), \tilde{x} - x \rangle$$

3) Assume that there exists  $M_1 \geq 0$  such that

$$\mathbb{E}[s(x; \xi)\|^2] \leq M_1^2.$$

Then, it follows that condition (A3) holds with  $\sigma = 2M_1$  and  $M = M_1$ . Indeed, for every  $x \in \text{dom } h$ ,

$$\begin{aligned} \|f'(x)\| &= \|\mathbb{E}[s(x, \xi)]\| \leq \mathbb{E}[\|s(x, \xi)\|] \\ &\leq (\mathbb{E}[\|s(x, \xi)\|^2])^{1/2} \leq M_1 = M \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[\|s(x, \xi) - f'(x)\|^2] &\leq \mathbb{E}[2\|f'(x)\|^2 + 2\|s(x, \xi)\|^2] \\ &\leq 2\|f'(x)\|^2 + 2\mathbb{E}[\|s(x, \xi)\|^2] \\ &\leq 2M_1^2 + 2M_1^2 = 4M_1^2 = \sigma^2 \end{aligned}$$

Two inequalities above are due to:

**Remark:** If  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and  $X(\xi)$  is a random variable, then

$$g(\mathbb{E}_\xi[X(\xi)]) \leq \mathbb{E}_\xi[g(X(\xi))]$$

or simply

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

## 2 Stochastic CS method

Let

$$\tilde{\ell}_f(\tilde{x}; x, s) = f(x) + \langle s, \tilde{x} - x \rangle.$$

### Stochastic composite subgradient method

(0) Let  $x_0 \in \text{dom } h$  be given, and set  $k = 0$  and

$$\lambda := \frac{\varepsilon}{(M^2 + \sigma^2) + \varepsilon L}; \quad (3)$$

(1) take a sample  $\xi_k$  of r.v.  $\xi$  which is independent from the previous samples  $\xi_0, \dots, \xi_{k-1}$  and set  $s_k = s(x_k, \xi_k)$ ;

(2) compute

$$x_{k+1} = \operatorname{argmin} \left\{ \tilde{\phi}_k^\lambda(x) := \tilde{\ell}_f(x; x_k, s_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 \right\} \quad (4)$$

where  $\tilde{\ell}_f(x; x_k, s_k) = f(x_k) + \langle s_k, x - x_k \rangle$ ;

(2) set  $k \leftarrow k + 1$  and go to step 1.

Let  $\mathbb{E}_{x_k}[\cdot]$  denote expectation of  $[\cdot]$  conditioned on  $x_k$ .

#### Remark:

1) relation (14) implies that

$$\frac{(M^2 + \sigma^2)\lambda}{1 - \lambda L} = \varepsilon \quad (5)$$

2) In view of (A2), have:

$$\mathbb{E}_{x_k}[s_k] = f'(x_k)$$

and hence

$$\mathbb{E}_{x_k}[\tilde{\ell}_f(x; x_k, s_k)] = \ell_f(x; x_k) \quad \forall x \in \mathbb{R}^n \quad (6)$$

Moreover,

$$\mathbb{E}_{x_k}[\|s_k - f'(x_k)\|^2] \leq \sigma^2 \quad (7)$$

### 3 Complexity analysis

**Lemma 3.1** *For every  $k \geq 0$  and  $x \in \text{dom } h$ , we have*

$$E[\tilde{\phi}_k^\lambda(x_{k+1})] - \phi(x) \leq \frac{1}{2\lambda} d_k^2(x) - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) d_{k+1}^2(x).$$

where

$$d_k^2(x) := E(\|x - x_k\|^2) \quad \forall k \geq 0.$$

**Proof:** Since the function  $\tilde{\phi}_k^\lambda$  defined in (15) is  $(\lambda^{-1} + \mu)$ -strongly convex, it follows from (15) that

$$\begin{aligned} \tilde{\phi}_k^\lambda(x_{k+1}) &\leq \tilde{\phi}_k^\lambda(x) - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) \|x - x_{k+1}\|^2 \\ &= \tilde{\ell}_f(x; x_k, s_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) \|x - x_{k+1}\|^2. \end{aligned}$$

Taking expectation conditioned on  $x_k$ , using (6), the definition of  $\phi$  in (1) and the fact that  $\ell_f(\cdot; x_k) \leq f(\cdot)$ , have

$$\begin{aligned} E_{x_k}[\tilde{\phi}_k^\lambda(x_{k+1})] &\leq \ell_f(x; x_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) E_{x_k}(\|x - x_{k+1}\|^2) \\ &\leq \phi(x) + \frac{1}{2\lambda} \|x - x_k\|^2 - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) E_{x_k}(\|x - x_{k+1}\|^2). \end{aligned}$$

The lemma follows by taking expectation of the above inequality.  $\blacksquare$

**Lemma 3.2** *For every  $k \geq 0$ , we have*

$$E \left[ \tilde{\phi}_k^\lambda(x_{k+1}) \right] \geq E[\phi(x_{k+1})] - \frac{\varepsilon}{2}$$

**Proof:** Using (2) and the definitions of  $\phi$  and  $\tilde{\phi}_k^\lambda$  in (1) and (15), respectively, we conclude that for every  $x \in \text{dom } h$ ,

$$\begin{aligned} \tilde{\phi}_k^\lambda(x) &= \tilde{\ell}_f(x; x_k, s_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 \\ &= \ell_f(x; x_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 + \langle s_k - f'(x_k), x - x_k \rangle \\ &\geq \left( f(x) - \frac{L}{2} \|x - x_k\|^2 - M \|x - x_k\| \right) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 + \langle s_k - f'(x_k), x - x_k \rangle \\ &\geq \phi(x) + \frac{1}{2} \left( \frac{1}{\lambda} - L \right) \|x - x_k\|^2 - (M + \|s_k - f'(x_k)\|) \|x - x_k\| \\ &\geq \phi(x) + \min_{t \in \mathbb{R}} \left\{ \frac{1}{2} \left( \frac{1}{\lambda} - L \right) t^2 - (M + \|s_k - f'(x_k)\|) t \right\} \\ &= \phi(x) - \frac{(M + \|s_k - f'(x_k)\|)^2 \lambda}{2(1 - \lambda L)} \geq \phi(x) - \frac{(M^2 + \|s_k - f'(x_k)\|^2) \lambda}{1 - \lambda L}. \end{aligned}$$

Taking  $x = x_{k+1}$ , have

$$\tilde{\phi}_k^\lambda(x_{k+1}) \geq \phi(x_{k+1}) - \frac{(M^2 + \|s_k - f'(x_k)\|^2) \lambda}{1 - \lambda L}.$$

Taking expectation conditioned on  $x_k$ , and using (5) and (7), have

$$E_{x_k} \left[ \tilde{\phi}_k^\lambda(x_{k+1}) \right] \geq E_{x_k}[\phi(x_{k+1})] - \frac{(M^2 + \sigma^2) \lambda}{1 - \lambda L} = E_{x_k}[\phi(x_{k+1})] - \frac{\varepsilon}{2}.$$

The lemma follows by taking expectation of the above inequality.  $\blacksquare$

**Lemma 3.3** *For every  $k \geq 0$  and  $x \in \text{dom } h$ , we have*

$$E[\phi(x_{k+1})] - \phi(x) \leq \frac{1}{2\lambda} d_k^2(x) - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) d_{k+1}^2(x) + \frac{\varepsilon}{2}$$

**Proof:** This result follows immediately from Lemmas 3.2 and 3.1. ■

Consider the case where  $\mu = 0$  and define

$$\bar{x}_k := \frac{1}{k} \sum_{i=1}^k x_i$$

Then, Lemma 3.3 implies that

$$\begin{aligned} \phi(x) + \frac{\varepsilon}{2} + \frac{d_0^2(x) - d_k^2(x)}{2k\lambda} &\geq \frac{1}{k} \sum_{i=1}^k \mathbb{E}[\phi(x_i)] = \mathbb{E} \left[ \frac{1}{k} \sum_{i=1}^k \phi(x_i) \right] \\ &\geq \mathbb{E} \left[ \phi \left( \frac{1}{k} \sum_{i=1}^k x_i \right) \right] = \mathbb{E}[\phi(\bar{x}_k)] \end{aligned}$$

Taking  $x = \text{Proj}_{X_*}(x_0)$ , have

$$\mathbb{E}[\phi(\bar{x}_k) - \phi_*] \leq \frac{\varepsilon}{2} + \frac{d_0^2}{2k\lambda}$$

**Final complexity:**

$$\mathcal{O}_1 \left( \frac{d_0^2}{\lambda\varepsilon} \right) = \mathcal{O}_1 \left( d_0^2 \left[ \frac{M^2 + \sigma^2}{\varepsilon^2} + \frac{L}{\varepsilon} \right] \right).$$

Consider now the case where  $\mu > 0$ . Recall that

$$E[\phi(x_k)] - \phi(x) \leq \frac{1}{2\lambda} d_{k-1}^2(x) - \frac{1}{2} \left( \frac{1}{\lambda} + \mu \right) d_k^2(x) + \frac{\varepsilon}{2} \quad \forall k \geq 1 \quad (*)$$

**Lemma 3.4** *Let  $\{t_k\}$  and  $\{\alpha_k\}$  be sequence of nonnegative scalars satisfying*

$$t_k \leq \alpha_{k-1} - \theta \alpha_k + \delta \quad \forall k \geq 1. \quad (8)$$

*for some positive scalars  $\theta$  and  $\delta > 0$ , and define*

$$T_k := \frac{1}{\Theta_k} \sum_{i=1}^k \theta^{i-1} t_i, \quad \Theta_k := \sum_{i=1}^k \theta^{i-1} \quad (9)$$

*Then, for every  $k \geq 0$ ,*

$$T_k \leq \frac{\alpha_0 - \theta^k \alpha_k}{\Theta_k} + \delta$$

**Proof:** Have

$$\begin{aligned} \Theta_k T_k &= \sum_{i=1}^k \theta^{i-1} t_i \leq \sum_{i=1}^k \theta^{i-1} (\alpha_{i-1} - \theta \alpha_i + \delta) = \sum_{i=1}^k (\theta^{i-1} \alpha_{i-1} - \theta^i \alpha_i + \theta^{i-1} \delta) \\ &= \alpha_0 - \theta^k \alpha_k + \sum_{i=1}^k \theta^{i-1} \delta \leq \alpha_0 - \theta^k \alpha_k + \Theta_k \bar{\delta}. \end{aligned}$$

■

**Lemma 3.5** *Let  $\{\bar{x}_k\} \subset \text{dom } h$  be a random sequence such that for every  $k \geq 1$ ,*

$$E[\phi(\bar{x}_k)] \leq \frac{1}{\Theta_k} \sum_{i=1}^k \theta^{i-1} E[\phi(x_i)] \quad (10)$$

*where  $\Theta_k$  is as in (9) with  $\theta := (1 + \mu\lambda)$ . Then, for every  $k \geq 0$  and  $x \in \text{dom } h$ , we have*

$$E[\phi(\bar{x}_k)] - \phi(x) \leq \frac{1}{2\Theta_k \lambda} (d_0^2(x) - \theta^k d_k^2(x)) + \frac{\varepsilon}{2}$$

*and*

$$\Theta_k \geq \max \{k, \theta^{k-1}\}.$$

**Proof:** First note that  $(*)$  implies that (8) holds with

$$\theta = (1 + \lambda\mu), \quad \delta = \frac{(M^2 + \sigma^2)\lambda}{(1 - \lambda L)}, \quad \alpha_k = \frac{1}{2\lambda} d_k^2(x), \quad t_k = E[\phi(x_k)] - \phi(x),$$

It then follows from (10) and the first conclusion of Lemma 3.4 that

$$[E[\phi(\bar{x}_k)] - \phi(x)] \leq \frac{1}{\Theta_k} \sum_{i=1}^k \theta^{i-1} [E[\phi(x_i)] - \phi(x)] \leq \frac{1}{2\Theta_k \lambda} (d_0^2(x) - \theta^k d_k^2(x)) + \frac{\varepsilon}{2},$$

and hence that the lemma holds. ■

Can take

$$\bar{x}_k := \frac{1}{\Theta_k} \sum_{i=1}^k \theta^{i-1} x_i$$

Taking  $x = Proj_{X_*}(x_0)$ , have

$$\mathbb{E}[\phi(\bar{x}_k) - \phi_*] \leq \frac{\varepsilon}{2} + \frac{d_0^2}{2\Theta_k\lambda} \leq \frac{\varepsilon}{2} + \frac{d_0^2}{2\theta^{k-1}\lambda}$$

**Final complexity:**

$$\begin{aligned} \mathcal{O}_1\left(\frac{1}{\theta-1} \log_1^+\left(\frac{d_0^2}{\lambda\varepsilon}\right)\right) &= \mathcal{O}_1\left(\left(1 + \frac{1}{\mu\lambda}\right) \log_1^+\left(\frac{d_0^2}{\lambda\varepsilon}\right)\right) \\ &= \mathcal{O}_1\left(\left[\left[1 + \frac{M^2 + \sigma^2}{\varepsilon\mu} + \frac{L}{\mu}\right] \log_1^+\left(d_0^2 \left[\frac{M^2 + \sigma^2}{\varepsilon^2} + \frac{L}{\varepsilon}\right]\right)\right]\right). \end{aligned}$$



Let  $w$  be a  $\nu$ -distance generating function with respect to  $\text{dom } h$

**Stochastic Prox Mirror method**

(0) Let  $x_0 \in \text{dom } h$  be given, and set  $k = 0$  and

$$\lambda := \frac{\varepsilon}{4(M^2 + \sigma^2) + \varepsilon L}; \quad (11)$$

(1) take a sample  $\xi_k$  of r.v.  $\xi$  which is independent from the previous samples  $\xi_0, \dots, \xi_{k-1}$  and set  $s_k = s(x_k, \xi_k)$ ;

(2) compute

$$x_{k+1} = \operatorname{argmin} \left\{ \tilde{\phi}_k^\lambda(x) := \tilde{\ell}_f(x; x_k, s_k) + h(x) + \frac{1}{\lambda} dw(x; x_k) \right\} \quad (12)$$

where  $\tilde{\ell}_f(x; x_k, s_k) = f(x_k) + \langle s_k, x - x_k \rangle$ ;

(2) set  $k \leftarrow k + 1$  and go to step 1.

Up to the constant  $\nu$  which also shows up in the complexity, the final iteration-complexity bound is similar to the one above

## 4 Application

Consider the finite sum problem (FSP)

$$\phi_* := \inf\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\}$$

where

$$f(x) := \frac{1}{m} \sum_{i=1}^m f_i(x) \quad \forall x \in \mathbb{R}^n$$

and the following assumptions hold:

- $h \in \overline{\text{Conv}}(\mathbb{R}^n)$
- for every  $i = 1, \dots, m$ , function  $f_i \in \overline{\text{Conv}}(\mathbb{R}^n)$  and  $\text{dom } f_i \supset \text{dom } h$ ;
- for every  $i = 1, \dots, m$ , there exists a function  $f'_i : \text{dom } h \rightarrow \mathbb{R}^n$  satisfying the following properties:
  - $f'_i(x) \in \partial f_i(x)$  for all  $x \in \text{dom } h$
  - there exists  $M_i \geq 0$  such that for every  $x \in \text{dom } h$ ,

$$\|f'_i(x)\|_* \leq M_i \tag{13}$$

- optimal solution set  $X_*$  is nonempty, and hence  $\phi_* \in \mathbb{R}$

**Stochastic composite subgradient method for FSP**

(0) Let  $x_0 \in \text{dom } h$  be given, and set  $k = 0$  and

$$\lambda := \frac{\varepsilon}{20\tilde{M}^2}, \quad \tilde{M} = \sqrt{\frac{1}{m} \sum_{i=1}^m M_i^2}; \quad (14)$$

(1) choose  $\xi_k \in \{1, \dots, m\}$  randomly with uniform distribution and set

$$s_k = f'_{\xi_k}(x_k)$$

(2) compute

$$x_{k+1} = \operatorname{argmin} \left\{ \tilde{\ell}_f(x; x_k, s_k) + h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 \right\} \quad (15)$$

where

$$\tilde{\ell}_f(x; x_k, s_k) = f(x_k) + \langle s_k, x - x_k \rangle$$

(2) set  $k \leftarrow k + 1$  and go to step 1.

Let  $\xi$  denote the random variable which takes value  $i$  with probability  $1/m$  for every  $i \in \{1, \dots, m\}$ . Define

$$s(x; \xi) := f'_\xi(x) \quad \forall x \in \text{dom } h$$

Function  $s(\cdot; \cdot)$  is a stochastic subgradient since for every  $x \in \mathbb{R}^n$ , have

$$\mathbb{E}_\xi[s(x; \xi)] = \mathbb{E}_\xi[f'_\xi(x)] = \frac{1}{m} \sum_{i=1}^m f'_i(x) = \frac{1}{m} \sum_{i=1}^m \partial f_i(x) = \partial f(x)$$

Clearly, the above method is the stochastic composite SM with the above subgradient oracle, i.e,

$$s_k = s(x_k; \xi_k) \quad \forall k \geq 0$$

Note that

$$\mathbb{E}[\|s(x; \xi)\|_*^2] = \mathbb{E}_\xi[\|f'_\xi(x)\|_*^2] = \frac{1}{m} \sum_{i=1}^m \|f'_i(x)\|_*^2 \leq \frac{1}{m} \sum_{i=1}^m M_i^2 = \tilde{M}^2$$

**Complexity of the stochastic SM:**

$$\mathcal{O}\left(\frac{d_0^2 \tilde{M}^2}{\varepsilon^2}\right)$$

**Complexity of the deterministic SM:**

$$\mathcal{O}\left(\frac{d_0^2 M^2}{\varepsilon^2}\right)$$

where  $M$  is a constant such that

$$M^2 \geq \sup\{\|f'(x)\|_*^2 : x \in \text{dom } h\}$$

e.g.,

$$\|f'(x)\|_*^2 = \left\| \frac{1}{m} \sum_{i=1}^m f'_i(x) \right\|_*^2 = \frac{1}{m^2} \left( \sum_{i=1}^m \|f'_i(x)\|_* \right)^2 \leq \frac{1}{m^2} \left( \sum_{i=1}^m M_i \right)^2 =: M^2$$

It is easy to see that

$$\frac{\tilde{M}^2}{M^2} \in [1, m]$$

So, the stochastic performs more iterations in general but evaluates less subgradients per iteration (one versus  $m$ )

**Example:** Consider

$$f_i(x) = |\langle a_i, x \rangle + b_i| \quad \forall i = 1, \dots, m$$

Then

$$\partial f_i(x) = \begin{cases} a_i & \text{if } \langle a_i, x \rangle + b_i > 0 \\ -a_i & \text{if } \langle a_i, x \rangle + b_i < 0 \\ [-a_i, a_i] & \text{if } \langle a_i, x \rangle + b_i = 0 \end{cases}$$

Then

$$M_i = \|a_i\|_2$$

and

$$\tilde{M}^2 = \frac{1}{m} \sum_{i=1}^m \|a_i\|^2 = \frac{1}{m} \|A\|_F^2 = \frac{1}{m} \sum_{i=1}^m \lambda_i(AA^T)$$

where

$$A = [a_1 \cdots a_m] \in \mathbb{R}^{n \times m}$$

On the other hand every  $s \in \partial f(x)$  is of the form

$$s = \frac{1}{m} \sum_{i=1}^m p_i a_i$$

where  $p_i \in [0, 1]$  for every  $i = 1, \dots, m$ . Hence

$$\|s\| = \left\| \frac{1}{m} Ap \right\| \leq \frac{1}{m} \|Ap\| \leq \frac{1}{m} \|A\| \|p\| \leq \frac{1}{\sqrt{m}} \|A\|$$

where  $p = (p_1, \dots, p_m)^T$ . So, we can take  $M$  such that

$$M^2 = \frac{1}{m} \|A\|^2 = \frac{1}{m} \lambda_{\max}(A^T A)$$

So,

$$\frac{\tilde{M}^2}{M^2} = \frac{\sum_{i=1}^m \lambda_i(A^T A)}{\lambda_{\max}(A^T A)} \in [1, m]$$

## 5 Randomized block coordinate (RBC) methods

### 5.1 Unaccelerated RBC Method

#### 5.1.1 Problem and assumptions

Consider the multiblock composite optimization sum (MCO) problem

$$\phi_* := \inf\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\}$$

where

$$x = (x_1, \dots, x_b) \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_b}, \quad h(x) := \sum_{i=1}^m h_i(x_i) \quad \forall x \in \mathbb{R}^n$$

and the following assumptions hold:

- $h_i \in \overline{\text{Conv}}(\mathbb{R}^{n_i})$  for every  $i = 1, \dots, b$
- for every  $i = 1, \dots, m$ , function  $f \in \overline{\text{Conv}}(\mathbb{R}^n)$  and for every  $i = 1, \dots, b$ , there exists  $L_i \geq 0$  such that

$$f(x + U_i d_i) \leq f(x) + \langle \nabla_i f(x), d_i \rangle + \frac{L_i}{2} \|d_i\|^2 \quad \forall x \in \mathbb{R}^n, d_i \in \mathbb{R}^{n_i}$$

where  $\nabla_i f(x)$  is the  $i$ -th block of  $\nabla f(x)$  and  $U_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^n$  is the linear map

$$U_i(d_i) = (0, \dots, d_i, 0, \dots, 0)$$

- optimal solution set  $X_*$  is nonempty, and hence  $\phi_* \in \mathbb{R}$

**Unaccelerated Random Block Coordinate (U-RBC) Method**

(0) Let  $x^0 = (x_1^0, \dots, x_b^0) \in \text{dom } h$  be given, and set  $k = 0$

(1) choose  $\xi_k \in \{1, \dots, m\}$  randomly according to the distribution  $p = (p_1, \dots, p_b) \in \Delta_{++}^{(b)}$  and set

$$x^{k+1} = x^k[\xi_k]$$

where

$$x^k[i] := x^k + U_i(\hat{x}_i^k - x_i^k) \quad \forall i = 1, \dots, b$$

and

$$\hat{x}_i^k = \underset{u_i}{\text{argmin}} \left\{ \langle \nabla_i f(x^k), u_i - x_i^k \rangle + h_i(u_i) + \frac{L_i}{2} \|u_i - x_i^k\|^2 \right\}$$

(2) set  $k \leftarrow k + 1$  and go to step 1.

**Remark:** The definition of  $x^k[i]$  implies that its  $j$ -th block is equal to  $x_j^k$  if  $j \neq i$  and is equal to  $\hat{x}_i^k$  if  $j = i$

**Notation:** For  $x = (x_1, \dots, x_b) \in \mathbb{R}^n$  and  $\eta = (\eta_1, \dots, \eta_b) \in \mathbb{R}_{++}^b$ , define

$$\|x\|_\eta^2 = \sum_{i=1}^b \eta_i \|x_i\|^2$$

**Proposition 5.1** *If  $p_i = 1/b$  for every  $i = 1, \dots, b$ , then*

$$\mathbb{E}[\phi(x^k) - \phi_*] \leq \frac{b}{b+k} [\phi(x^0) - \phi_* + \|x^0 - x_*\|_L^2] \quad \forall x_* \in X_*$$

where  $L = (L_1, \dots, L_b)$

Hence the  $\varepsilon$ -iteration complexity of U-RBC is  $\mathcal{O}(b/\varepsilon)$ , more specifically,

$$\mathcal{O}_1 \left( b \left[ \frac{\phi(x^0) - \phi_* + \|x^0 - x_*\|_L^2}{\varepsilon} \right] \right)$$

### 5.1.2 Analysis

Let

$$\hat{x}^k = (\hat{x}_1^k, \dots, \hat{x}_b^k)$$

Then

$$\begin{aligned}\hat{x}^k &= \operatorname{argmin}_u \left\{ \langle \nabla f(x^k), u - x^k \rangle + h(u) + \frac{1}{2} \|u - x^k\|_L^2 \right\} \\ &= \operatorname{argmin}_u \left\{ \ell_f(u; x^k) + \frac{1}{2} \|u - x^k\|_L^2 \right\}\end{aligned}$$

In the lemmas below,  $x$ ,  $x^+$  and  $\hat{x}$  denotes  $x^k$ ,  $x^{k+1}$  and  $\hat{x}^k$ , respectively.

**Lemma 5.2** *For every  $u \in \operatorname{dom} h$ , have*

$$\ell_f(u; x) + h(u) + \frac{1}{2} \|u - x\|_L^2 \geq \ell_f(\hat{x}; x) + h(\hat{x}) + \frac{1}{2} \|\hat{x} - x\|_L^2 + \frac{1}{2} \|u - \hat{x}\|_L^2$$

**Proof:** This follows from the convex analysis result that I mentioned at the beginning of the course. ■

**Lemma 5.3** *For every  $u \in \operatorname{dom} h$ , have*

$$\frac{1}{2} \|u - x\|_L^2 - \frac{1}{2} \|u - \hat{x}\|_L^2 \geq \left[ \ell_f(\hat{x}; x) + h(\hat{x}) + \frac{1}{2} \|\hat{x} - x\|_L^2 \right] - \phi(u)$$

**Proof:** Follows from the previous lemma and the fact that  $\phi(\cdot) \geq \ell_f(\cdot; x) + h(\cdot)$ . ■

**Lemma 5.4** *Have*

$$\ell_f(\hat{x}; x) + h(\hat{x}) + \frac{1}{2} \|\hat{x} - x\|_L^2 \geq \phi(x) + \sum_{i=1}^b [\phi(x[i]) - \phi(x)]$$

**Proof:** Have

$$\begin{aligned}\ell_f(\hat{x}; x) + \frac{1}{2} \|\hat{x} - x\|_L^2 &= f(x) + \sum_{i=1}^b \left[ \langle \nabla_i f(x), \hat{x}_i - x_i \rangle + \frac{L_i}{2} \|\hat{x}_i - x_i\|^2 \right] \\ &\geq f(x) + \sum_{i=1}^b [f(x[i]) - f(x)]\end{aligned}$$

Also,

$$\begin{aligned}h(\hat{x}) &= h(x) + [h(\hat{x}) - h(x)] = h(x) + \sum_{i=1}^b [h_i(\hat{x}_i) - h_i(x_i)] \\ &= h(x) + \sum_{i=1}^b [h(x[i]) - h(x)]\end{aligned}$$

The conclusion of the lemma now follows by summing the above two inequalities and using the fact that  $\phi = f + h$ . ■



**Lemma 5.5** For every  $u \in \text{dom } h$ , have

$$\frac{1}{2}\|u - x\|_L^2 - \frac{1}{2}\|u - \hat{x}\|_L^2 \geq \left( \sum_{i=1}^b [\phi(x[i]) - \phi(x)] \right) + \phi(x) - \phi(u)$$

**Proof:** Follows trivially from Lemmas 5.3 and 5.4 ■

**Lemma 5.6** For every  $\eta = (\eta_1, \dots, \eta_b) \in \mathbb{R}_{++}^b$  and  $u \in \mathbb{R}^n$ , have

$$\mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] - \|x - u\|_\eta^2 = \|\hat{x} - u\|_{p\eta}^2 - \|x - u\|_{p\eta}^2$$

where  $\xi \in \{1, \dots, b\}$  is a random variable with distribution  $p = (p_1, \dots, p_b)$ .

**Proof:** Let  $u \in \mathbb{R}^n$  be given. First note that for any  $\alpha \in \mathbb{R}_{++}^b$  and  $i \in \{1, \dots, b\}$ , have

$$\|x[i] - u\|_\alpha^2 - \|x - u\|_\alpha^2 = \alpha_i (\|\hat{x}_i - u_i\|^2 - \|x_i - u_i\|^2)$$

Hence

$$\begin{aligned} \mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] - \|x - u\|_\eta^2 &= \left( \sum_{i=1}^b p_i \|x[i] - u\|_\eta^2 \right) - \|x - u\|_\eta^2 \\ &= \sum_{i=1}^b p_i \left( \|x[i] - u\|_\eta^2 - \|x - u\|_\eta^2 \right) \\ &= \sum_{i=1}^b p_i \eta_i \left( \|\hat{x}_i - u_i\|^2 - \|x_i - u_i\|^2 \right) \\ &= \|\hat{x} - u\|_{\eta p}^2 - \|x - u\|_{\eta p}^2 \end{aligned}$$

where the third equality is due to the first observation above and the last one is due to the definition of  $\|\cdot\|_\alpha$ . ■

**Lemma 5.7** Assume that  $\xi \in \{1, \dots, b\}$  is a random variable with distribution  $p = (p_1, \dots, p_b)$ . Then, for every  $u \in \text{dom } h$ , have

$$\|x - u\|_\eta^2 - \mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] \geq \phi(x) - \phi(u) + \frac{1}{p_{\min}} (\mathbb{E} [\phi(x[\xi]) - \phi(x)])$$

where  $\eta := L/p$  and  $L = (L_1, \dots, L_b)$ .

**Proof:** It follows from Lemma 5.5 and Lemma 5.22 with  $\eta = L/p$  that

$$\|x - u\|_\eta^2 - \mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] \geq \left( \sum_{i=1}^b [\phi(x[i]) - \phi(x)] \right) + \phi(x) - \phi(u)$$

Now, using the fact that  $\phi(x[i]) \leq \phi(x)$  for every  $i = 1, \dots, b$ , we have

$$p_{\min} \sum_{i=1}^b [\phi(x[i]) - \phi(x)] \geq \sum_{i=1}^b p_i [\phi(x[i]) - \phi(x)] = \mathbb{E} [\phi(x[\xi]) - \phi(x)]$$

The conclusion of the lemma now follows by combining the above two relations. ■

Translated to the context of U-RBC, the result above implies the following result.

**Lemma 5.8** *Let  $\eta = L/p$ . then for every  $u \in \mathbb{R}^n$ , have*

$$\frac{1}{2} \|x^k - u\|_\eta^2 - \frac{1}{2} \mathbb{E}_{x^k} [\|x^{k+1} - u\|_\eta^2] \geq \phi(x^k) - \phi(u) + \frac{1}{p_{\min}} \mathbb{E}_{x^k} [\phi(x^{k+1}) - \phi(x^k)]$$

**Lemma 5.9** *Let  $\eta = L/p$  and define*

$$d_{k,\eta}(u) := \left( \mathbb{E} [\|x^k - u\|_\eta^2] \right)^{1/2}, \quad \theta_k(u) := \mathbb{E} [\phi(x^k) - \phi(u)].$$

*Then, for every  $u \in \mathbb{R}^n$ , have*

$$\frac{1}{2} [d_{k,\eta}(u)]^2 - \frac{1}{2} [d_{k+1,\eta}(u)]^2 \geq \theta_k(u) + \frac{1}{p_{\min}} [\theta_{k+1}(u) - \theta_k(u)]$$

**Proposition 5.10** *If  $\eta = L/p$  then, for every  $u \in \mathbb{R}^n$ ,*

$$\theta_k(u) \leq \frac{1}{(1/p_{\min}) + k} \left[ \frac{1}{2} [d_{0,\eta}(u)]^2 - \frac{1}{2} [d_{k,\eta}(u)]^2 + \frac{1}{p_{\min}} \theta_0(u) \right].$$

**Proof:** Summing the inequality in the previous lemma from  $k = 0$  to  $k = k - 1$ , have

$$\frac{1}{p_{\min}} [\theta_0(u) - \theta_k(u)] + \frac{1}{2} [d_{0,\eta}(u)]^2 - \frac{1}{2} [d_{k,\eta}(u)]^2 = \sum_{i=0}^{k-1} \theta_i(u) \geq k \theta_k(u)$$

since  $\theta_{k-1}(u) \geq \theta_k(u)$  for every  $k \geq 1$ . The conclusion of the proposition immediately follows from the above inequality. ■

Let us now express the above bound in terms of

$$d_0 = \text{dist}(x_0; X_*)$$

where the distance is in terms of  $\|\cdot\|$ .

**Corollary 5.11** *For every  $u \in \mathbb{R}^n$ ,*

$$\theta_k(u) \leq \frac{1}{(1/p_{\min}) + k} \left[ \frac{1}{2} \max \left( \frac{L}{p} \right) \|x_0 - u\|^2 + \frac{1}{p_{\min}} [\phi(x^0) - \phi(u)] \right].$$

**Proof:** Have

$$\begin{aligned} [d_{0,\eta}(u)]^2 &= \|x^0 - u\|_\eta^2 = \sum_{i=1}^b \eta_i \|x_i^0 - u_i\|^2 = \sum_{i=1}^b \frac{L_i}{p_i} \|x_i^0 - u_i\|^2 \\ &\leq \max \left( \frac{L}{p} \right) \sum_{i=1}^b \|x_i^0 - u_i\|^2 = \|x_0 - u\|^2 \max \left( \frac{L}{p} \right) \end{aligned}$$

The result now follows from the above proposition and the definition of  $\theta_0(u)$ . ■

**Corollary 5.12** For every  $u \in \mathbb{R}^n$ ,

$$\mathbb{E} [\phi(x^k) - \phi_*] \leq \frac{1}{(1/p_{\min}) + k} \left[ \frac{1}{2} \max \left( \frac{L}{p} \right) d_0^2 + \frac{1}{p_{\min}} [\phi(x^0) - \phi_*] \right].$$

**Proof:** Follows from the previous corollary with  $u = Proj_{X_*}(x_0)$  ■

**Remarks:**

1) The above bound can be further refined to

$$\mathbb{E} [\phi(x^k) - \phi_*] \leq \frac{1}{1 + kp_{\min}} \left[ \frac{L_{\max} d_0^2}{2} + [\phi(x^0) - \phi_*] \right].$$

2) If  $p_i = 1/b$  for every  $i = 1, \dots, b$  then

$$\mathbb{E} [\phi(x^k) - \phi_*] \leq \frac{b}{b + k} \left[ \frac{L_{\max} d_0^2}{2} + [\phi(x^0) - \phi_*] \right].$$

**Question:** What is the Lipschitz constant of  $\nabla f(\cdot)$  in terms of the  $L_i$ 's?

## 5.2 Accelerated RBC Method

### 5.2.1 The method

We start by stating the method

#### A-RBC Method

(0) Let  $x^0 = (x_1^0, \dots, x_b^0) \in \text{dom } h$  be given, and set  $k = 0$ ,  $y^0 = x^0$ ,  $A_0 = 0$

(1) compute  $a_k > 0$  such that

$$\frac{a_k^2}{A_k + a_k} = \frac{1}{b^2},$$

and set

$$A_{k+1} = A_k + a_k, \quad \tilde{x}^k = \frac{A_k y^k + a_k x^k}{A_{k+1}}$$

(2) choose  $\xi_k \in \{1, \dots, m\}$  randomly according to the uniform distribution  $p = (1/b, \dots, 1/b) \in \Delta_{++}^{(b)}$ , and compute

$$x^{k+1} = x^k[\xi_k], \quad y^{k+1} = \tilde{x}_k + \frac{1}{ba_k}(x^{k+1} - x^k) = \tilde{x}_k + \frac{ba_k}{A_{k+1}}(x^{k+1} - x^k)$$

where

$$x^k[i] = x^k + U_i(\hat{x}_i^k - x_i^k),$$

$$\hat{x}_i^k = \operatorname{argmin}_u \left( a_k [\langle \nabla_i f(\tilde{x}_i^k), u - \tilde{x}_i^k \rangle + h_i(u)] + \frac{L_i}{2b} \|u - x_i^k\|^2 \right).$$

(3) set  $k \leftarrow k + 1$  and go to step 1.

**Proposition 5.13** *Assume that  $\eta = L/p$  and  $p = (1/b, \dots, 1/b)$ . Then, for every  $k \geq 0$ , the following statements hold:*

a) have

$$\begin{aligned} A_{k+1} \mathbb{E} [\phi(y^{k+1}) - \phi(u)] + \frac{1}{2b} \mathbb{E} [\|x^{k+1} - u\|_\eta^2] \\ \leq A_k \mathbb{E} [\phi(y^k) - \phi(u)] + \frac{1}{2b} \mathbb{E} [\|x^k - u\|_\eta^2] \end{aligned}$$

b)  $A_k \geq k^2/4b^2$

**Corollary 5.14** *Assume that  $\eta = L/p$  and  $p = (1/b, \dots, 1/b)$ . Then, for every  $k \geq 0$ ,*

$$\mathbb{E} [\phi(y^k) - \phi(u)] \leq \frac{2b}{k^2} \|x^0 - u\|_\eta^2 \leq \frac{2b^2}{k^2} \|x^0 - u\|_L^2$$

*As a consequence,*

$$\mathbb{E} [\phi(y^k) - \phi_*] \leq \frac{2b^2}{k^2} \max(L) d_0^2$$

Hence, the  $\varepsilon$ -iteration complexity of A-RBC is  $\mathcal{O}(b/\sqrt{\varepsilon})$ , more specifically,

$$\mathcal{O}_1 \left( bd_0 \sqrt{\frac{\max(L)}{\varepsilon}} \right)$$

### 5.2.2 Analysis

**Lemma 5.15** *Have*

$$\hat{x} = \operatorname{argmin}_u \left( a\gamma(u) + \frac{1}{2b} \|u - x\|_L^2 \right).$$

where

$$\gamma(\cdot) := \ell_f(\cdot; \hat{x}) + h(\cdot)$$

**Proof:** Obvious. ■

**Lemma 5.16** *For every  $u \in \operatorname{dom} h$ , have*

$$a\gamma(\hat{x}) + \frac{1}{2b} \|\hat{x} - x\|_L^2 \leq a\gamma(u) + \frac{1}{2b} \|u - x\|_L^2 - \frac{1}{2b} \|u - x\|_L^2.$$

**Proof:** This follows from the convex analysis result that I mentioned at the beginning of the course. ■

**Lemma 5.17** *For every  $i = 1, \dots, b$ , define*

$$y[i] = \tilde{x} + \frac{1}{ba} (x[i] - x) = \tilde{x} + \frac{ba}{A^+} (x[i] - x).$$

*Then, for every  $i = 1, \dots, b$ , have*

$$y[i] = \frac{A}{A^+} y + \frac{a}{A^+} x_b[i]$$

where

$$x_b[i] = x + bU_i(\hat{x}_i - x_i) = x + b(x[i] - x).$$

**Proof:** We have

$$\begin{aligned} y[i] &= \tilde{x} + \frac{ba}{A^+} (x[i] - x) \\ &= \frac{Ay + ax}{A^+} + \frac{ba}{A^+} (x[i] - x) \\ &= \frac{A}{A^+} + \frac{a}{A^+} \left( \underbrace{x + b(x[i] - x)}_{x_b[i]} \right). \end{aligned}$$
■

**Lemma 5.18** *We have  $\frac{1}{b} \sum_{i=1}^b x_b[i] = \hat{x}$ .*

**Proof:** We have

$$x_b[i] = x + b(x[i] - x) = x + bU_i(\hat{x}_i - x_i)$$

and hence

$$\frac{1}{b} \sum_{i=1}^b x_b[i] = x + \frac{1}{b} [b(\hat{x} - x)] = \hat{x}.$$
■

**Lemma 5.19** *If  $h$  is an indicator function and  $y[i] \in \text{dom } h$ , then*

$$A^+ \phi(y[i]) \leq A\phi(y) + a\tilde{\gamma}(x_b[i]) + \frac{L_i}{2} \|\hat{x}_i - x_i\|^2$$

where  $\tilde{\gamma}(u) = \ell_f(u; \tilde{x})$ .

**Proof:** We have

$$\phi(y[i]) \leq \ell_f(y[i]; \tilde{x}) + h(y[i]) + \frac{L_i}{2} \|y[i] - \tilde{x}\|^2 = \tilde{\gamma}(y[i]) + \frac{L_i}{2} \|y[i] - \tilde{x}\|^2$$

So,

$$\begin{aligned} A^+ \phi(y[i]) &\leq A^+ \left[ \tilde{\gamma}(y[i]) + \frac{L_i}{2} \|y[i] - \tilde{x}\|^2 \right] \\ \text{(by Lemma (5.17))} \quad &= A^+ \left[ \tilde{\gamma} \left( \frac{A}{A^+} y + \frac{a}{A^+} x_b[i] \right) + \frac{L_i b^2 a^2}{2(A^+)^2} \|x[i] - x\|^2 \right] \\ \text{(convexity of } \tilde{\gamma}) \quad &\leq A\tilde{\gamma}(y) + a\tilde{\gamma}(x_b[i]) + \frac{L_i}{2} \|x[i] - x\|^2 \\ &\leq A\phi(y) + a\tilde{\gamma}(x_b[i]) + \frac{L_i}{2} \|x[i] - x\|^2 \end{aligned}$$

■

**Lemma 5.20** *If  $h$  is an indicator function, have*

$$A^+ \mathbb{E} [\phi(y[\xi])] \leq A\phi(y) + a\gamma(\hat{x}) + \frac{1}{2b} \|\hat{x} - x\|_L^2.$$

**Proof:**

$$\begin{aligned} A^+ \mathbb{E} [\phi(y[\xi])] &= A^+ \left( \frac{1}{b} \sum_{i=1}^b \phi(y[i]) \right) \\ &\leq A\phi(y) + a \left( \frac{1}{b} \sum_{i=1}^b \tilde{\gamma}(x_b[i]) \right) + \frac{1}{2b} \|\hat{x} - x\|_L^2 \\ \text{(by Lemma 5.18)} \quad &\leq A\phi(y) + a\tilde{\gamma}(\hat{x}) + \frac{1}{2b} \|\hat{x} - x\|_L^2 \\ &\leq A\phi(y) + a\gamma(\hat{x}) + \frac{1}{2b} \|\hat{x} - x\|_L^2 \end{aligned}$$

■

**Lemma 5.21** *If  $h$  is an indicator function, we have*

$$A^+ \mathbb{E} [\phi(y[\xi]) - \phi(u)] \leq A[\phi(y) - \phi(u)] + \frac{1}{2b} \|u - x\|_L^2 - \frac{1}{2b} \|u - \hat{x}\|_L^2.$$

**Proof:** Have

$$\begin{aligned} A^+ \mathbb{E} [\phi(y[\xi])] &\leq A\phi(y) + a\gamma(u) + \frac{1}{2b} \|u - x\|_L^2 - \frac{1}{2b} \|u - \hat{x}\|_L^2 \\ &\leq A\phi(y) + a\phi(u) + \frac{1}{2b} \|u - x\|_L^2 - \frac{1}{2b} \|u - \hat{x}\|_L^2. \end{aligned}$$

The conclusion of the lemma now follows by subtracting  $A^+ \phi(u)$  from both sides and using the fact that  $A^+ = A + a$ .  $\blacksquare$

**Lemma 5.22** *For every  $\eta = (\eta_1, \dots, \eta_b) \in \mathbb{R}_{++}^b$  and  $u \in \mathbb{R}^n$ , have*

$$\mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] - \|x - u\|_\eta^2 = \|\hat{x} - u\|_{p\eta}^2 - \|x - u\|_{p\eta}^2$$

where  $\xi \in \{1, \dots, b\}$  is a random variable with distribution  $p = (p_1, \dots, p_b)$ .

Hence, if  $\eta = bL$  and  $p_i = 1/b$  for every  $i = 1, \dots, b$ , then

$$\|x - u\|_\eta^2 - \mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] = \|x - u\|_L^2 - \|\hat{x} - u\|_L^2$$

**Lemma 5.23** *Assume that  $h$  is an indicator function and let  $\eta = L/p$ . Then, for all  $u \in \mathbb{R}^n$ ,*

$$A^+ \mathbb{E} [\phi(y[\xi]) - \phi(u)] + \frac{1}{2b} \mathbb{E} \left[ \|x[\xi] - u\|_\eta^2 \right] \leq A[\phi(y) - \phi(u)] + \frac{1}{2b} \|x - u\|_\eta^2$$