Randomized Block Coordinate Descent

Lecturer: Jiaming Liang *November 9, 2023*

1 Motivation

In this lecture, we discuss methods for solving optimization problems with huge-scale and block-wise decomposible structure. Since first-order methods with full gradient updates would be computationally expensive, we are interested in methods that make partial gradient/vector updates, i.e., an update in only one block of the full gradient/vector. Methods of this type are called coordinate descent methods.

1.1 Theoretical justification

The simplest variant of the coordinate descent method is based on a cyclic coordinate search. However, for this strategy it is difficult to prove convergence, and almost impossible to estimate the rate of convergence

Another possibility is to move along the direction corresponding to the component of gradient with maximal absolute value. Consider

$$\min_{x \in \mathbb{R}^n} f(x)$$

where the convex objective function f has component-wise Lipschitz continuous gradient, i.e.,

$$|\nabla_i f(x + he_i) - \nabla_i f(x)| \leq M|h|, \quad x \in \mathbb{R}^n,\ h \in \mathbb{R},\ i = 1, \ldots, n.$$

Consider the following algorithm.

Algorithm 1 Maximum abosolute value coordinate descent

Input: Initial point $x_0 \in \mathbb{R}^n$
for $k \geq 0$ **do**
Step 1. Choose

$$i_k = \operatorname{argmax}_{1 \leq i \leq n} |\nabla_i f(x_k)|$$

Step 2. Update

$$x_{k+1} = x_k - \frac{1}{M} \nabla_{i_k} f(x_k)\, e_{i_k}.$$

end for

It is not difficult to show that

$$
\begin{aligned}
f\left(x_k\right)-f\left(x_{k+1}\right) &\geq \frac{1}{2M}\left|\nabla_{i_k} f\left(x_k\right)\right|^2 \geq \frac{1}{2nM}\left\|\nabla f\left(x_k\right)\right\|^2 \\
&\geq \frac{1}{2nMR^2}\left(f\left(x_k\right)-f_*\right)^2
\end{aligned}
$$

where $R \geq \left\|x_0-x^*\right\|$, and hence that

$$
f\left(x_k\right)-f_* \leq \frac{2nMR^2}{k+4}, \quad k \geq 0.
$$

Since the maximum absolute value coordinate is needed, this method still requires computation of the full gradient. However, if this vector is available, it seems better to apply the usual full gradient methods. It is also important that for convex functions with Lipschitz-continuous gradient, i.e.,

$$
\|\nabla f(x)-\nabla f(y)\| \leq L\|x-y\|, \quad x, y \in \mathbb{R}^n
$$

it can happen that $M \geq L$.

1.2 Computational complexity

In huge-scale optimization, the computation of full gradient or directional derivative evaluations is expansive, and even a function value can require substantial computational efforts. Moreover, some parts of the problem's data can be distributed in space and in time. The problem's data may be only partially available at the moment of evaluating the current test point.

Example.

$$
\min_{x \in \mathbb{R}^n}\left\{f(x):=\sum_{i=1}^n f_i\left(x^{(i)}\right)+\frac{1}{2}\|Ax-b\|^2\right\}
$$

where f_i are convex differentiable univariate functions, $A=\left(a_1, \ldots, a_n\right) \in \mathbb{R}^{p \times n}$, and $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^p$. Then

$$
\begin{aligned}
\nabla_i f(x) &= f_i^{\prime}\left(x^{(i)}\right)+\left\langle a_i, g(x)\right\rangle, \quad i=1, \ldots, n \\
g(x) &= Ax-b.
\end{aligned}
$$

If the residual vector $g(x)$ is already computed, then the computation of i-th directional derivative requires $O\left(p_i\right)$ operations, where p_i is the number of nonzero elements in vector a_i. On the other hand, the coordinate move $x_+=x+\alpha e_i$ results in the following change in the residual:

$$
g\left(x_+\right)=g(x)+\alpha a_i.
$$

Therefore, the i-th coordinate step for problem (1.6) needs $O\left(p_i\right)$ operations. Note that computation of either the function value, or the whole gradient, or an arbitrary directional derivative requires $O\left(\sum_{i=1}^n p_i\right)$ operations.

2 Randomized block coordinate descent

Define the partition of the identity matrix

$$I_n = (U_1, \cdots, U_b) \in \mathbb{R}^{n\times n}, \quad U_i \in \mathbb{R}^{n\times n_i}, \quad i = 1, \cdots, b.$$

Thus, any $x = (x^{(1)}, \cdots, x^{(b)})^T \in \mathbb{R}^n$ can be represented as

$$x = \sum_{i=1}^{b} U_i x^{(i)}, \quad x^{(i)} \in \mathbb{R}^{n_i}, \quad i = 1, \cdots, b.$$

Consider the problem of minimizing a composite convex function:

$$\min_{x\in\mathbb{R}^n} \{\phi(x) := f(x) + h(x)\}.$$

Assumptions for f and h:

- h is closed, convex, and separable, i.e., $h(x) = \sum_{i=1}^{b} h_i(x^{(i)})$;
- f is convex and differentiable on $\operatorname{dom} h$ and there exists $L_i \geq 0$ for $i = 1, \cdots, b$ such that

$$f(x + U_i(x'^{(i)} - x^{(i)})) - [f(x) + \langle \nabla_i f(x), x'^{(i)} - x^{(i)}\rangle] \leq \frac{L_i}{2}\|x'^{(i)} - x^{(i)}\|^2 \quad \forall x, x' \in \operatorname{dom} h. \quad (1)$$

Define the randomized block coordinate descent update as follows:

$$x^{(i)} = \operatorname{argmin}_{u^{(i)}\in\mathbb{R}^{n_i}} \left(\langle \nabla_i f(x_0), u^{(i)} - x_0^{(i)}\rangle + h_i(u^{(i)}) + \frac{1}{2\lambda_i}\left\|u^{(i)} - x_0^{(i)}\right\|^2 \right), \quad (2)$$

and

$$x[i] = x_0 + U_i(x^{(i)} - x_0^{(i)}), \quad i = 1, \cdots, b. \quad (3)$$

2.1 The method

Now, we are ready to state the randomized block coordinate descent method.

Algorithm 2 Randomized block coordinate descent

Input: Initial point $x_0 \in \operatorname{dom} h$

for $k \geq 1$ **do**

Step 1. Generate a random variable i_k according to

$$\mathbf{Prob}(i_k = i) = p_i, \quad i = 1, 2, \cdots, b.$$

Step 2. Compute x_k by the randomized block coordinate descent update (2) and (3).

end for

Lemma 1. *Applying the block coordinate descent update in the i-th coordinate, we have*

$$\phi(x_0) - \phi(x[i]) \geq \varepsilon_i + \lambda_i \left(1 - \frac{L_i \lambda_i}{2}\right) \|r^{(i)}\|^2, \tag{4}$$

where

$$r^{(i)} = \frac{x_0^{(i)} - x^{(i)}}{\lambda_i}, \quad \varepsilon_i := \varepsilon(x[i]) = h_i(x_0^{(i)}) - h_i(x^{(i)}) - \langle r^{(i)} - \nabla_i f(x_0), x_0^{(i)} - x^{(i)} \rangle.$$

Note: $\varepsilon_i = \varepsilon(x[i])$ is a random variable, and $r^{(i)} \in \mathbb{R}^{n_i}$ is a random vector. They both depend on the choice of i-th coordinate.

Proof. First note in the i-th block, we have the following equalities to connect the local and global quantities

$$\langle \nabla_i f(x_0), x_0^{(i)} - x^{(i)} \rangle = \langle \nabla f(x_0), x_0 - x[i] \rangle \tag{5}$$

$$h_i(x_0^{(i)}) - h_i(x^{(i)}) = h(x_0) - h(x[i]) \tag{6}$$

$$\|x[i] - x_0\|^2 = \|x_0^{(i)} - x^{(i)}\|^2 = \lambda_i^2 \|r^{(i)}\|^2 \tag{7}$$

The optimality condition for (2) is

$$r^{(i)} \in \nabla_i f(x_0) + \partial h_i(x^{(i)}), \quad \text{or } r^{(i)} \in \nabla_i f(x_0) + \partial_{\varepsilon_i} h_i(x_0^{(i)}),$$

where $\varepsilon_i = h_i(x_0^{(i)}) - h_i(x^{(i)}) - \langle r^{(i)} - \nabla_i f(x_0), x_0^{(i)} - x^{(i)} \rangle$.
Hence

$$\varepsilon_i = h_i(x_0^{(i)}) - h_i(x^{(i)}) + \langle \nabla_i f(x_0), x_0^{(i)} - x^{(i)} \rangle - \lambda_i \|r^{(i)}\|^2,$$

and

$$\varepsilon_i + \lambda_i \|r^{(i)}\|^2 = h_i(x_0^{(i)}) - h_i(x^{(i)}) + \langle \nabla_i f(x_0), x_0^{(i)} - x^{(i)} \rangle.$$

It follows from (5) and (6) that

$$\begin{aligned}
\varepsilon_i + \lambda_i \|r^{(i)}\|^2 &= h(x_0) - h(x[i]) + \langle \nabla f(x_0), x_0 - x[i] \rangle \\
&= (f + h)(x_0) - h(x[i]) - (f(x_0) + \langle \nabla f(x_0), x[i] - x_0 \rangle) \\
&= (f + h)(x_0) - h(x[i]) - \ell_f(x[i]; x_0) \\
&\leq (f + h)(x_0) - h(x[i]) - \left(f(x[i]) - \frac{L_i}{2} \|x[i] - x_0\|^2\right),
\end{aligned}$$

where the last inequality is due to (1). Then by (7), we have

$$\varepsilon_i + \lambda_i \|r^{(i)}\|^2 \leq (f + h)(x_0) - h(x[i]) - \left(f(x[i]) - \frac{L_i \lambda_i^2}{2} \|r^{(i)}\|^2\right),$$

so

$$\varepsilon_i + \lambda_i \left(1 - \frac{L_i \lambda_i}{2}\right) \|r^{(i)}\|^2 \leq (f + h)(x_0) - (f + h)(x[i]).$$

□

Definition 1. *Define*

$$\|r\|_{\#}^2 = \sum_{i=1}^{b} \frac{p_i \lambda_i}{2} \|r^{(i)}\|^2,$$

and

$$(\|s\|_{\#}^*)^2 = \sum_{i=1}^{b} (\frac{p_i \lambda_i}{2})^{-1} (\|s^{(i)}\|^*)^2 = \sum_{i=1}^{b} (\frac{p_i \lambda_i}{2})^{-1} \|s^{(i)}\|^2.$$

Lemma 2. *Chooseing* $\lambda_i = \frac{1}{L_i}$, $i = 1, \cdots, b$, *and applying the randomized block coordinate descent, we have*

$$\phi(x_0) - \mathbb{E}[\phi(x)] - \sum_{i=1}^{b} p_i \varepsilon_i \geq \|r\|_{\#}^2.$$

Proof. Taking expectation on both sides of (4),

$$\begin{aligned}
\phi(x_0) - \mathbb{E}[\phi(x)] &= \sum_{i=1}^{b} p_i \left(\phi(x_0) - \phi(x[i])\right) \\
&\geq \sum_{i=1}^{b} p_i \left(\varepsilon_i + \lambda_i \left(1 - \frac{L_i \lambda_i}{2}\right) \|r^{(i)}\|^2\right) = \sum_{i=1}^{b} p_i \left(\varepsilon_i + \frac{\lambda_i}{2} \|r^{(i)}\|^2\right).
\end{aligned}$$

It follows from the above inequality and Definition 1 that

$$\phi(x_0) - \mathbb{E}[\phi(x)] \geq \sum_{i=1}^{b} p_i \varepsilon_i + \sum_{i=1}^{b} \frac{p_i \lambda_i}{2} \|r^{(i)}\|^2 = \sum_{i=1}^{b} p_i \varepsilon_i + \|r\|_{\#}^2.$$

□

Lemma 3.

$$\|r\|_{\#} \geq \frac{\phi(x_0) - \phi(x_*) - \sum_{i=1}^{b} \varepsilon_i}{\|x_0 - x_*\|_{\#}^*}.$$

Proof. The optimality condition for (2) is

$$r^{(i)} \in \nabla_i f(x_0) + \partial h_i(x^{(i)}), \quad \text{or } r^{(i)} \in \nabla_i f(x_0) + \partial_{\varepsilon_i} h_i(x_0^{(i)}).$$

From the latter inclusion, we have

$$h_i(u^{(i)}) \geq h_i(x_0^{(i)}) + \langle r^{(i)} - \nabla_i f(x_0), u^{(i)} - x_0^{(i)} \rangle - \varepsilon_i, \quad \forall u^{(i)} \in \mathbb{R}^{n_i}.$$

Taking $u^{(i)} = x_*^{(i)}$, we have

$$h_i(x_*^{(i)}) \geq h_i(x_0^{(i)}) + \langle r^{(i)} - \nabla_i f(x_0), x_*^{(i)} - x_0^{(i)} \rangle - \varepsilon_i.$$

Summation over coordinates gives

$$h(x_*) \geq h(x_0) + \langle r - \nabla f(x_0), x_* - x_0 \rangle - \sum_{i=1}^{b} \varepsilon_i.$$

Thus, we have

$$\begin{aligned}\langle r, x_0 - x_* \rangle + \sum_{i=1}^{b} \varepsilon_i &\geq h(x_0) - h(x_*) - \langle \nabla f(x_0), x_* - x_0 \rangle \\ &\geq h(x_0) - h(x_*) + f(x_0) - f(x_*) \\ &= \phi(x_0) - \phi(x_*).\end{aligned}$$

Hence, by the above inequality and the Cauchy-Schwarz inequality, we obtain

$$\|r\|_{\#} \geq \frac{\phi(x_0) - \phi(x_*) - \sum_{i=1}^{b} \varepsilon_i}{\|x_0 - x_*\|_{\#}^*}.$$

□

Definition 2. *Define*

$$\xi_k = \{i_0, i_1, \cdots, i_k\}$$

to be the sequence of observed random variables after k iterations, where i_k is the choice of block in the k-th iteration.

Definition 3. *Define the expected values*

$$\phi_k = \mathbb{E}_{\xi_k}[\phi(x_k)], \quad \bar{\varepsilon}_k = \mathbb{E}_{\xi_k}[\varepsilon(x_k)],$$

and

$$\Delta_k = \phi_k - \phi_*, \quad \tau_k = \frac{1}{\Delta_k}.$$

Lemma 4.

$$\phi_k - \phi_{k+1} \geq \bar{\varepsilon}_{k+1},$$

and

$$\Delta_k - \Delta_{k+1} \geq \bar{\varepsilon}_{k+1}.$$

Proof. Given ξ_k, it follows from Lemma 2 that

$$\phi(x_k) - \mathbb{E}_{i_{k+1}}[\phi(x_{k+1})] \geq \sum_{i=1}^{b} p_i \varepsilon_i = \mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})].$$

Taking the expectation in ξ_k, we get

$$\phi_k - \phi_{k+1} \geq \bar{\varepsilon}_{k+1}.$$

Using Definition 3, we have

$$\Delta_k - \Delta_{k+1} \geq \bar{\varepsilon}_{k+1}.$$

□

2.2 Uniform distribution

In this subsection, we consider the uniform distribution of p_i's, where $p_i = \frac{1}{b}$, $i = 1, \cdots, b$. For an iteration $k \geq 1$,we discuss two cases: $\phi(x_k) - \phi_* \leq 2b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$ and $\phi(x_k) - \phi_* \geq 2b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$.

First, we note that in the uniform distribution case,

$$\phi(x_k) - \phi_* - b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})] = \phi(x_k) - \phi_* - b\sum_{i=1}^{b} p_i \varepsilon(x_{k+1}[i]) = \phi(x_k) - \phi_* - \sum_{i=1}^{b} \varepsilon_i$$

where $\varepsilon_i = \varepsilon(x_{k+1}[i])$.

Case 1. $\phi(x_k) - \phi_* \leq 2b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$

Taking the expectation in ξ_k, then we have

$$\frac{1}{2}(\mathbb{E}_{\xi_k}[\phi(x_k)] - \phi_*) \leq b\mathbb{E}_{\xi_{k+1}}[\varepsilon(x_{k+1})],$$

or equivalently,

$$\frac{1}{2}\Delta_k = \frac{1}{2}(\phi_k - \phi_*) \leq b\bar{\varepsilon}_{k+1}. \tag{8}$$

Proposition 1.

$$\Delta_{k+1} \leq C_1 \Delta_k,$$

and

$$\tau_{k+1} \geq \frac{1}{C_1}\tau_k,$$

where $C_1 = 1 - \frac{1}{2b} < 1$.

Proof. It follows from (8) and Lemma 4 that

$$\Delta_k - \Delta_{k+1} \geq \bar{\varepsilon}_{k+1} \geq \frac{1}{2b}\Delta_k.$$

Thus, we have

$$\Delta_{k+1} \leq \left(1 - \frac{1}{2b}\right)\Delta_k = C_1 \Delta_k,$$

and

$$\tau_k \leq C_1 \tau_{k+1}.$$

□

Case 2. $\phi(x_k) - \phi_* \geq 2b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$

Proposition 2.

$$\tau_{k+1} - \tau_k \geq \frac{1}{C},$$

where

$$C = 4R^2, \text{ and } R = \max_x \left\{ \max_{x_* \in X_*} \|x - x_*\|_{\#}^* : \phi(x) \leq \phi(x_0) \right\}$$

which is a measure of the size of the level set of ϕ given by x_0.
Note: in the uniform distribution case,

$$\|s\|_{\#}^* = \left(2b \sum_{i=1}^{b} L_i \|s^{(i)}\|^2 \right)^{1/2}.$$

Proof. By Lemmas 2 and 3, we have

$$\phi(x_0) - \mathbb{E}[\phi(x)] - \sum_{i=1}^{b} p_i \varepsilon_i \geq \|r\|_{\#}^2 \geq \frac{(\phi(x_0) - \phi_* - \sum_{i=1}^{b} \varepsilon_i)^2}{(\|x_0 - x_*\|_{\#}^*)^2}.$$

For the k-th iteration, that is

$$\begin{aligned} \phi(x_k) - \mathbb{E}_{i_{k+1}}[\phi(x_{k+1})] - \mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})] &\geq \frac{(\phi(x_k) - \phi_* - b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})])^2}{(\|x_k - x_*\|_{\#}^*)^2} \\ &\geq \frac{(\phi(x_k) - \phi_*)^2}{4(\|x_k - x_*\|_{\#}^*)^2} \geq \frac{(\phi(x_k) - \phi_*)^2}{C}, \end{aligned}$$

where the second inequality is due to the assumption that $\phi(x_k) - \phi_* \geq 2b\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$ and the last inequality is due to the definition of C.

Taking the expectation in ξ_k and using the Jensen's inequality, we obtain

$$\phi_k - \phi_{k+1} - \bar{\varepsilon}_{k+1} \geq \frac{\mathbb{E}_{\xi_k}(\phi(x_k) - \phi_*)^2}{C} \geq \frac{(\phi_k - \phi_*)^2}{C}.$$

Thus, we have

$$\Delta_k - \Delta_{k+1} \geq \Delta_k - \Delta_{k+1} - \bar{\varepsilon}_{k+1} \geq \frac{1}{C}(\Delta_k)^2,$$

and hence

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} = \frac{\Delta_k - \Delta_{k+1}}{\Delta_k \Delta_{k+1}} \geq \frac{\Delta_k - \Delta_{k+1}}{(\Delta_k)^2} \geq \frac{1}{C},$$

or equivalently,

$$\tau_{k+1} - \tau_k \geq \frac{1}{C}.$$

□

Definition 4. *Let*

$$K^+ := \{j : \frac{1}{2}\Delta_j \geq b\bar{\varepsilon}_{j+1}, 0 \leq j \leq k-1\}, \quad K^- := \{j : \frac{1}{2}\Delta_j \leq b\bar{\varepsilon}_{j+1}, 0 \leq j \leq k-1\}.$$

Theorem 1.

$$\Delta_k \leq \frac{\max\{4R^2, (2b-1)[\phi(x_0) - \phi_*]\}}{k}.$$

Proof. Using Propositions 1 and 2, we have

$$\begin{aligned} \tau_k - \tau_0 &= \sum_{j\in K^+} (\tau_j - \tau_{j-1}) + \sum_{j \in K^-} (\tau_j - \tau_{j-1}) \\ &\geq |K^+| \frac{1}{C} + |K^-| \tau_0 \left(\frac{1}{C_1} - 1 \right) \\ &\geq (|K^+| + |K^-|) \min \left\{ \frac{1}{C}, \tau_0 \left(\frac{1}{C_1} - 1 \right) \right\} \\ &= \frac{k}{C'}, \end{aligned}$$

where $C' = \max\{C, C_1/(\tau_0(1-C_1))\}$. Therefore, we have

$$\frac{1}{\Delta_k} = \tau_k \geq \tau_0 + \frac{k}{C'} \geq \frac{k}{C'},$$

and finally

$$\Delta_k \leq \frac{C'}{k}.$$

□

2.3 Arbitrary distribution

In this subsection, we consider the arbitrary distribution. W.L.O.G., we can assume $0 < p_1 \leq p_2 \leq \cdots \leq p_b < 1$, thus

$$p_1 \sum_{i=1}^{b} \varepsilon_i = \min_{1\leq i \leq b} p_i \sum_{i=1}^{b} \varepsilon_i \leq \sum_{i=1}^{b} p_i \varepsilon_i = \mathbb{E}[\varepsilon(x)].$$

For an iteration $k \geq 1$,we discuss two cases: $\phi(x_k) - \phi_* \leq \frac{2}{p_1}\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$ and $\phi(x_k) - \phi_* \geq \frac{2}{p_1}\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$.

We present the following results without giving their proofs since they are similar to those in Subsection 2.2.

Case 1. $\phi(x_k) - \phi_* \leq \frac{2}{p_1}\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$

Proposition 3.

$$\Delta_{k+1} \leq C_2 \Delta_k.$$

and

$$\tau_{k+1} \geq \frac{1}{C_2}\tau_k,$$

where $C_2 = 1 - \frac{p_1}{2} < 1$.

Case 2. $\phi(x_k) - \phi_* \geq \frac{2}{p_1}\mathbb{E}_{i_{k+1}}[\varepsilon(x_{k+1})]$

Proposition 4.

$$\tau_{k+1} - \tau_k \geq \frac{1}{C},$$

where

$$C = 4R^2, \text{ and } R = \max_x \left\{ \max_{x_* \in X_*} \|x - x_*\|_\#^* : \phi(x) \leq \phi(x_0) \right\}.$$

Theorem 2.

$$\Delta_k \leq \frac{\max\{4R^2, (2/p_1 - 1)[\phi(x_0) - \phi_*]\}}{k}.$$

Example. One choice of the non-uniform distribution is for some $\alpha \geq 0$,

$$p_i = \frac{L_i^\alpha}{S_\alpha},$$

where $S_\alpha = \sum_{i=1}^b L_i^\alpha$. In this case, if $\lambda_i = 1/L_i$, then

$$\|r\|_\# = \left(\sum_{i=1}^b \frac{L_i^{\alpha-1}}{2S_i} \|r^{(i)}\|^2 \right)^{1/2},$$

and

$$\|s\|_\#^* = \left(2S_\alpha \sum_{i=1}^b L_i^{1-\alpha} \|s^{(i)}\|^2 \right)^{1/2}.$$

3 Dual problem

In this section, we show that the dual of the regularized empirical risk minimization (ERM) problems associated with linear predictors is in the block-wise decomposible structure .

Let $A_1, A_2, \cdots, A_n$ be the columns of $A \in \mathbb{R}^{d \times n}$, $\phi_1, \phi_2, \cdots, \phi_n$ be a sequence of convex functions defined on $\mathbb{R}$, and g be a convex function defined on $\mathbb{R}^d$. The goal of regularized ERM with linear predictors is to solve the following convex optimization problem

$$\text{minimize}_{w \in \mathbb{R}^d} \left\{ P(w) := \frac{1}{n} \sum_{i=1}^n \phi_i(A_i^T w) + \lambda g(w) \right\}.$$

Reformulated primal problem

$$\text{minimize}_{y\in\mathbb{R}^n, w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n \phi_i(y_i) + \lambda g(w) : y_i = A_i^T w \right\}.$$

Dual function of the reformulated problem is

$$\begin{aligned}
&\inf_{y\in\mathbb{R}^n, w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \phi_i(y_i) + \lambda g(w) + \sum_{i=1}^n x_i(y_i - A_i^T w) \\
=& \sum_{i=1}^n \left[\inf_{y_i\in\mathbb{R}} \frac{1}{n}\phi_i(y_i) + x_i y_i \right] + \inf_{w\in\mathbb{R}^d} \lambda g(w) - (Ax)^T w \\
=& \sum_{i=1}^n \left[-\frac{1}{n} \sup_{y_i\in\mathbb{R}} y_i(-nx_i) - \phi_i(y_i) \right] - \lambda \sup_{w\in\mathbb{R}^d} w^T(\frac{1}{\lambda}Ax) - g(w) \\
=& \sum_{i=1}^n -\frac{1}{n}\phi_i^*(-nx_i) - \lambda g^*(\frac{1}{\lambda}Ax) \\
=& -\frac{1}{n}\sum_{i=1}^n \phi_i^*(-u_i) - \lambda g^*(\frac{1}{n\lambda}Au)
\end{aligned}$$

where $u = nx$.
The dual problem is

$$\text{maximize}_{x\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n -\phi_i^*(-u_i) - \lambda g^*(\frac{1}{\lambda n}Au). \right\}.$$

This is equivalent to minimizing

$$\text{minimize}_{x\in\mathbb{R}^n} \left\{ F(u) := \frac{1}{n}\sum_{i=1}^n \phi_i^*(-u_i) + \lambda g^*(\frac{1}{\lambda n}Au) \right\}.$$

The structure of $F(x)$ matches our general formulation of the composite convex function with

$$f(u) = \lambda g^*(\frac{1}{\lambda n}Au), \quad h(u) = \sum_{i=1}^n h_i(u_i) = \frac{1}{n}\sum_{i=1}^n \phi_i^*(-u_i).$$

4 Accelerated randomized block coordinate descent

In this section, we develop a variant of the randomized block coordinate descent method that achieves the acceleration convergence rate $\mathcal{O}(k^{-2})$. For simplicity, we consider the for uniform distribution.

Algorithm 3 Accelerated randomized block coordinate descent

Input: Initial point $x_0 \in \operatorname{dom} h$. Set $y_0 = x_0$ and $A_0 = 1 - \frac{1}{b}$.

for $k \geq 0$ **do**

Step 1. Generate a random variable $i_{k+1} = i$ uniformly from $\{1, 2, \ldots, b\}$;

Step 2. Compute

$$a_k = \frac{1 + \sqrt{1 + 4b^2 A_k}}{2b^2}, \quad A_{k+1} = A_k + a_k, \quad \tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_k}{A_{k+1}} x_k$$

Step 3. Compute

$$x_{k+1}^{(i)} := \operatorname{argmin}_{u^{(i)}} \{\langle \nabla_i f(\tilde{x}_k), u^{(i)} - \tilde{x}_k^{(i)} \rangle + h_i(u^{(i)}) + \frac{L_i}{2a_k b} \|u^{(i)} - x_k^{(i)}\|^2\}, \tag{9}$$

$$\begin{aligned} x_{k+1} &= x_k + U_i(x_{k+1}^{(i)} - x_k^{(i)}), \\ y_{k+1}^{(i)} &:= \tilde{x}_k^{(i)} + \frac{1}{a_k b}(x_{k+1}^{(i)} - x_k^{(i)}), \\ y_{k+1} &= \tilde{x}_k + U_i(y_{k+1}^{(i)} - \tilde{x}_k^{(i)}). \end{aligned} \tag{10}$$

end for

We first make some basic observations.

Lemma 5.

$$A_{k+1} = a_k^2 b^2, \quad A_k \geq \frac{k^2}{4b^2}.$$

Proof. The identity follows from the facts that $A_{k+1} = A_k + a_k$ and a_k is the solution of

$$b^2 a_k^2 - a_k - A_k = 0.$$

Now, we prove the inequality. It follows from the definition of a_k that

$$a_k = \frac{1 + \sqrt{1 + 4b^2 A_k}}{2b^2} \geq \frac{1}{2b^2} + \frac{\sqrt{A_k}}{b},$$

thus

$$A_{k+1} = A_k + a_k \geq A_k + \frac{\sqrt{A_k}}{b} + \frac{1}{2b^2} \geq \left(\sqrt{A_k} + \frac{1}{2b}\right)^2.$$

Hence

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2b}.$$

Summing this equality over iterations gives

$$\sqrt{A_k} \geq \sqrt{A_0} + \frac{k}{2b} \geq \frac{k}{2b}.$$

This concludes the proof. ∎

Lemma 6. *Define* $\beta_1^0 = A_1 - a_0 b = 0$, $\beta_1^1 = a_0 b = 1$ *and for* $k \geq 1$,

$$\beta_{k+1}^l = \begin{cases} \beta_k^l, & l = 0, \cdots, k-1; \\ \beta_k^k + a_k - a_k b, & l = k; \\ a_k b & l = k+1. \end{cases} \tag{11}$$

Then, for all $k \geq 1$, *we have*

$$\beta_k \geq 0, \quad l = 0, 1, \ldots, k,$$

and

$$\sum_{l=0}^{k} \beta_k^l = A_k, \quad A_k y_k = \sum_{l=0}^{k} \beta_k^l x_l, \tag{12}$$

That is, y_k *is a convex combination of* $x_0, x_1, \cdots, x_k$.

Proof. Since β_1^0 and β_1^1 are nonnegative. Using an induction argument and (11), it amounts to prove

$$\beta_{k+1}^k = \beta_k^k + a_k - a_k b \geq 0.$$

It follows from $\beta_k^k = a_{k-1} b$ that

$$\begin{aligned} \beta_{k+1}^k &= a_{k-1} b + a_k - a_k b \\ &= \frac{1}{a_k + a_{k-1}} \left[a_k (a_k + a_{k-1}) - (a_k^2 - a_{k-1}^2) b \right] \\ &= \frac{1}{a_k + a_{k-1}} \left[a_k (a_k + a_{k-1}) - (A_{k+1} - A_k) \frac{1}{b} \right] \\ &= \frac{a_k}{a_k + a_{k-1}} \left(a_k + a_{k-1} - \frac{1}{b} \right). \end{aligned}$$

Since $a_0 = \frac{1}{b}$ and $\{a_k\}$ is increasing, we have $a_k + a_{k-1} \geq a_1 + a_0 \geq 2a_0 \geq \frac{2}{b}$ and hence $\beta_{k+1}^k \geq 0$.

We now prove (12) by induction. First, it easy to check that (12) holds for $k = 1$. Assume that (12) holds for some $k \geq 1$. Then, it follows from (11) and the induction hypothesis that

$$\begin{aligned} \sum_{l=0}^{k+1} \beta_{k+1}^l &= \sum_{l=0}^{k-1} \beta_{k+1}^l + \beta_{k+1}^k + \beta_{k+1}^{k+1} \\ &= \sum_{l=0}^{k-1} \beta_k^l + a_k = \sum_{l=0}^{k} \beta_k^l + a_k \\ &= A_k + a_k = A_{k+1.} \end{aligned}$$

Moreover, using (11) and the induction hypothesis, we also have

$$
\begin{aligned}
A_{k+1}y_{k+1} &= A_{k+1}\left(\tilde{x}_k + \frac{1}{a_k b}(x_{k+1}-x_k)\right) \\
&= A_k y_k + a_k x_k + a_k b(x_{k+1}-x_k) \\
&= \sum_{l=0}^{k} \beta_k^l x_l + a_k x_k + a_k b(x_{k+1}-x_k) \\
&= \sum_{l=0}^{k-1} \beta_k^l x_l + \beta_k^k x_k + a_k x_k + a_k b(x_{k+1}-x_k) \\
&= \sum_{l=0}^{k-1} \beta_k^l x_l + (\beta_k^k + a_k - a_k b)x_k + a_k b x_{k+1} \\
&= \sum_{l=0}^{k-1} \beta_{k+1}^l x_l + \beta_{k+1}^k x_k + \beta_{k+1}^{k+1} x_{k+1} \\
&= \sum_{l=0}^{k+1} \beta_{k+1}^l x_l.
\end{aligned}
$$

□

Lemma 7. *Define* $\Gamma_0 = h(x_0)$ *and for* $k \geq 1$,

$$\Gamma_k = \frac{\sum_{l=0}^k \beta_k^l h(x_l)}{A_k}.$$

Then, we have

$$\Gamma_k \geq h(y_k), \quad A_{k+1}\Gamma_{k+1} = A_k\Gamma_k + a_k(1-b)h(x_k) + a_k b h(x_{k+1}).$$

Proof. Using the definition of Γ_k, the second equality in (12), and the convexity of h, we have

$$A_k\Gamma_k = \sum_{l=0}^{k} \beta_k^l h(x_l) \geq A_k h\left(\frac{\sum_{l=0}^k \beta_k^l x_l}{A_k}\right) = A_k h(y_k).$$

Hence, the inequality holds. Now, we show the identity. Using the definitions of β_{k+1}^l in (11), we have

$$\begin{aligned}
A_{k+1}\Gamma_{k+1} &= \sum_{l=0}^{k+1} \beta_{k+1}^l h(x_l) = \sum_{l=0}^{k-1} \beta_{k+1}^l h(x_l) + \beta_{k+1}^k h(x_k) + \beta_{k+1}^{k+1} h(x_{k+1}) \\
&= \sum_{l=0}^{k-1} \beta_k^l h(x_l) + [\beta_k^k + a_k(1-b)]h(x_k) + a_k b h(x_{k+1}) \\
&= \sum_{l=0}^{k} \beta_k^l h(x_l) + a_k(1-b)h(x_k) + a_k b h(x_{k+1}) \\
&= A_k\Gamma_k + a_k(1-b)h(x_k) + a_k b h(x_{k+1}).
\end{aligned}$$

□

Lemma 8. *Assuming $i_{k+1} = i$, then we have*

$$\begin{aligned}
&A_{k+1}[\Gamma_{k+1} + f(y_{k+1})] \\
\leq &A_k[\Gamma_k + f(y_k)] + a_k\left[(1-b)h(x_k) + bh(x_{k+1})\right] + a_k\ell_f(x_b[i];\tilde{x}_k) + \frac{L_i}{2}\|x_{k+1} - x_k\|^2, \qquad (13)
\end{aligned}$$

where

$$x_b[i] = x_k + bU_i\left(x_{k+1}^{(i)} - x_k^{(i)}\right) = x_k + b(x_{k+1} - x_k). \qquad (14)$$

Proof. For simplicity, we will omit the iteration index k and let

$$x^+[i] = x_{k+1}, \quad y^+[i] = y_{k+1}, \quad x_+^{(i)} = x_{k+1}^{(i)}, \quad y_+^{(i)} = y_{k+1}^{(i)}.$$

Observations from step 3 of Algorithm 3

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

and

$$y^+[i] = \frac{A}{A^+}y + \frac{a}{A^+}x_b[i]. \qquad (15)$$

Indeed,

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

It follows from the smoothness of f (see (1)) and (10), we have

$$\begin{aligned}
f(y^+[i]) &\leq f(\tilde{x}) + \langle \nabla_i f(\tilde{x}), y_+^{(i)} - \tilde{x}^{(i)} \rangle + \frac{L_i}{2}\|y_+^{(i)} - \tilde{x}^{(i)}\|^2 \\
&= \ell_f(y^+[i];\tilde{x}) + \frac{L_i}{2}\|y^+[i] - \tilde{x}\|^2.
\end{aligned}$$

Using this inequality, Lemma 1, and (15), we obtain

$$
\begin{aligned}
&A^+[\Gamma^+ + f(y^+[i])] - a\left[(1-b)h(x) + bh(x^+[i])\right] \\
=&A\Gamma + A^+ f(y^+[i]) \\
\le&A\Gamma + A^+\left(\ell_f(y^+[i];\tilde{x}) + \frac{L_i}{2}\|y^+[i]-\tilde{x}\|^2\right) \\
=&A\Gamma + A^+\left[\ell_f\left(\frac{Ay + ax_b[i]}{A^+};\tilde{x}\right) + \frac{L_i}{2(ab)^2}\|x^+[i]-x\|^2\right] \\
=&A\Gamma + A\ell_f(y;\tilde{x}) + a\ell_f(x_b[i];\tilde{x}) + \frac{L_i}{2}\|x^+[i]-x\|^2 \\
\le&A[\Gamma + f(y)] + a\ell_f(x_b[i];\tilde{x}) + \frac{L_i}{2}\|x^+[i]-x\|^2,
\end{aligned}
$$

where the last inequality follows from the convexity of f. □

Lemma 9. *Define*

$$
\hat{x}_{k+1} = \left(x_{k+1}^{(1)}, x_{k+1}^{(2)}, \cdots, x_{k+1}^{(b)}\right), \quad \|r\|_L = \left(\sum_{i=1}^{b} L_i\|r^{(i)}\|^2\right)^{1/2}.
$$

Then, the following statements hold

(i)

$$
\hat{x}_{k+1} = \operatorname{argmin}_{u\in\mathbb{R}^n}\left\{a_k\left[\ell_f(u;\tilde{x}_k) + h(u)\right] + \frac{1}{2b}\|u - x_k\|_L^2\right\};
$$

(ii) for any $u \in \mathbb{R}^n$,

$$
\|x_{k+1} - u\|_L^2 - \|x_k - u\|_L^2 = L_i\|x_{k+1}^{(i)} - u^{(i)}\|^2 - L_i\|x_k^{(i)} - u^{(i)}\|^2;
$$

(iii) for any $u \in \mathbb{R}^n$, under total expectation,

$$
\mathbb{E}[\|x_{k+1} - u\|_L^2 - \|x_k - u\|_L^2] = \frac{1}{b}\mathbb{E}[\|\hat{x}_{k+1} - u\|_L^2] - \frac{1}{b}\mathbb{E}[\|x_k - u\|_L^2];
$$

(iv) let x_b denotes $x_b[i]$ when $i_{k+1} = i$, we have

$$
\mathbb{E}_{i_{k+1}}[x_b] = \hat{x}_{k+1}, \quad \mathbb{E}_{i_{k+1}}[(1-b)h(x_k) + bh(x_{k+1})] = h(\hat{x}_{k+1}).
$$

Proof. (i) This statement immediately follows from (9) and the definition of $\hat{x}_{k+1}$.

(ii) and (iii) directly follow from the definitions of $\hat{x}_{k+1}$ and $\|\cdot\|_L$.

(iv) Assuming $i_{k+1} = i$, then the first identity directly follows from the definition of $x_b[i]$ in (14). Using the fact that $h(x) = \sum_{i=1}^b h_i(x^{(i)})$, we have

$$\mathbb{E}_i[h(x^+[i])] = \frac{1}{b}\sum_{i=1}^b h(x^+[i]) = \frac{1}{b}\sum_{i=1}^b \left(\sum_{j\neq i} h_j(x^{(j)}) + h_i(x_+^{(i)})\right) = \frac{b-1}{b}h(x) + \frac{1}{b}h(\hat{x}^+).$$

Rearranging the above equation and using the definition of $x_b[i]$ in (14), we have

$$\mathbb{E}_i[bh(x^+[i]) - (b-1)h(x)] = h(\hat{x}).$$

□

Theorem 3.

$$\mathbb{E}_{\xi_k}[\phi(y_k)] - \phi^* \le \frac{4(b^2-b)(\phi(y_0)-\phi^*) + 2b^2\|x_0 - x_*\|_L^2}{k^2}.$$

Proof. Taking the expectation of (13) in i_{k+1} and using Lemma 9, we have

$$A_{k+1}\mathbb{E}_{i_{k+1}}[\Gamma_{k+1} + f(y_{k+1})] \le A_k[\Gamma_k + f(y_k)] + a_k h(\hat{x}_{k+1}) + a_k \ell_f(\hat{x}_{k+1}; \tilde{x}_k) + \frac{1}{2b}\|\hat{x}_{k+1} - x_k\|_L^2.$$

It follows form Lemma 9(i) that for any $u \in \operatorname{dom} h$

$$a_k[h(\hat{x}_{k+1}) + \ell_f(\hat{x}_{k+1};\tilde{x}_k)] + \frac{1}{2b}\|\hat{x}_{k+1} - x_k\|_L^2 \le a_k[h(u) + \ell_f(u;\tilde{x}_k)] + \frac{1}{2b}\|u - x_k\|_L^2 - \frac{1}{2b}\|u - \hat{x}_{k+1}\|_L^2.$$

Combing the above inequalities, we obtain

$$\begin{aligned} A_{k+1}\mathbb{E}_{i_{k+1}}[\Gamma_{k+1} + f(y_{k+1})] &\le A_k[\Gamma_k + f(y_k)] + a_k h(u) + a_k \ell_f(u;\tilde{x}_k) + \frac{1}{2b}\|u - x_k\|_L^2 - \frac{1}{2b}\|u - \hat{x}_{k+1}\|_L^2 \\ &\le A_k[\Gamma_k + f(y_k)] + a_k\phi(u) + \frac{1}{2b}\|u - x_k\|_L^2 - \frac{1}{2b}\|u - \hat{x}_{k+1}\|_L^2, \end{aligned}$$

where the last inequality follows from the convexity of f and the fact that $\phi = f + h$. Taking $u = x_*$ in the above inequality, taking the expectation in ξ_k (i.e., total expectation), and , we have

$$A_{k+1}\mathbb{E}[\Gamma_{k+1} + f(y_{k+1})] \le A_k\mathbb{E}[\Gamma_k + f(y_k)] + a_k\phi_* + + \frac{1}{2b}\mathbb{E}[\|x_* - x_k\|_L^2] - \frac{1}{2b}\mathbb{E}[\|x_* - \hat{x}_{k+1}\|_L^2].$$

This inequality together with Lemma 9(iii) implies that

$$A_{k+1}\mathbb{E}[\Gamma_{k+1} + f(y_{k+1})] \le A_k\mathbb{E}[\Gamma_k + f(y_k)] + a_k\phi_* + \frac{1}{2}\mathbb{E}[\|x_* - x_k\|_L^2] - \frac{1}{2}\mathbb{E}[\|x_* - x_{k+1}\|_L^2].$$

Rearranging terms givs

$$A_{k+1}(\mathbb{E}[\Gamma_{k+1} + f(y_{k+1})] - \phi_*) + \frac{1}{2}\mathbb{E}[\|x_{k+1} - x_*\|_L^2] \le A_k(\mathbb{E}[\Gamma_k + f(y_k)] - \phi_*) + \frac{1}{2}\mathbb{E}[\|x_k - x_*\|_L^2].$$

It follows from Lemma 7 that

$$\begin{aligned}2A_k(\mathbb{E}[\phi(y_k)] - \phi_*) &\leq 2A_k(\mathbb{E}[\Gamma_k + f(y_k)] - \phi_*)\\ &\leq 2A_0([\Gamma_0 + f(y_0)] - \phi_*) + \|x_0 - x_*\|_L^2\\ &= 2A_0(\phi(y_0) - \phi_*) + \|x_0 - x_*\|_L^2.\end{aligned}$$

Finally, using Lemma 5 and $A_0 = 1 - 1/b$, we conclude that

$$\begin{aligned}\mathbb{E}[\phi(y_k)] - \phi^* &\leq \frac{2A_0(\phi(y_0) - \phi^*) + \|x_0 - x_*\|_L^2}{2A_k}\\ &\leq \frac{4(b^2 - b)(\phi(y_0) - \phi^*) + 2b^2\|x_0 - x_*\|_L^2}{k^2}.\end{aligned}$$

□

In the case where $b = 1$, we can recover the result of the standard ACG method, that is, Theorem 3 becomes the same as Theorem 1 of Lecture 7, i.e.,

$$\mathbb{E}_{\xi_k}[\phi(y_k)] - \phi^* \leq \frac{2L\|x_0 - x_*\|^2}{k^2}.$$