12. Coordinate descent methods

- theoretical justifications
- randomized coordinate descent method
- minimizing composite objectives
- accelerated coordinate descent method
consider smooth unconstrained minimization problem:

\[
\begin{align*}
\text{minimize} \quad & f(x) \\
x \in & \mathbb{R}^N
\end{align*}
\]

- coordinate blocks: \( x = (x_1, \ldots, x_n) \) with \( x_i \in \mathbb{R}^{N_i} \) and \( \sum_{i=1}^{n} N_i = N \)
- more generally, partition with a permutation matrix: \( U = [U_1 \cdots U_n] \)

\[
x_i = U_i^T x, \quad x = \sum_{i=1}^{n} U_i x_i
\]

- blocks of gradient:

\[
\nabla_i f(x) = U_i^T \nabla f(x)
\]

- coordinate update:

\[
x^+ = x - t U_i \nabla_i f(x)
\]
(Block) coordinate descent

choose $x^{(0)} \in \mathbb{R}^n$, and iterate for $k = 0, 1, 2, \ldots$

1. choose coordinate $i(k)$
2. update $x^{(k+1)} = x^{(k)} - t_k U_{i_k} \nabla_{i_k} f(x^{(k)})$

• among the first schemes for solving smooth unconstrained problems

• cyclic or round-Robin: difficult to analyze convergence

• mostly local convergence results for particular classes of problems

• does it really work (better than full gradient method)?
Steepest coordinate descent

choose \(x^{(0)} \in \mathbb{R}^n\), and iterate for \(k = 0, 1, 2, \ldots\)

1. choose \(i(k) = \arg\max_{i \in \{1, \ldots, n\}} \|\nabla_i f(x^{(k)})\|_2\)

2. update \(x^{(k+1)} = x^{(k)} - t_k U_{i(k)} \nabla_{i(k)} f(x^{(k)})\)

assumptions

- \(\nabla f(x)\) is block-wise Lipschitz continuous

\[
\|\nabla_i f(x + U_i v) - \nabla_i f(x)\|_2 \leq L_i \|v\|_2, \quad i = 1, \ldots, n
\]

- \(f\) has bounded sub-level set, in particular, define

\[
R(x) = \max_y \left\{ \max_{x^* \in X^*} \|y - x^*\|_2 : f(y) \leq f(x) \right\}
\]
Analysis for constant step size

quadratic upper bound due to block coordinate-wise Lipschitz assumption:

\[ f(x + U_i v) \leq f(x) + \langle \nabla_i f(x), v \rangle + \frac{L_i}{2} \| v \|_2^2, \quad i = 1, \ldots, n \]

assume constant step size \( 0 < t \leq 1/M \), with \( M \triangleq \max_{i \in \{1, \ldots, n\}} L_i \)

\[ f(x^+) \leq f(x) - \frac{t}{2} \| \nabla_i f(x) \|_2^2 \leq f(x) - \frac{t}{2n} \| \nabla f(x) \|_2^2 \]

by convexity and Cauchy-Schwarz inequality,

\[ f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle \]

\[ \leq \| \nabla f(x) \|_2 \| x - x^* \|_2 \leq \| \nabla f(x) \|_2 R(x^{(0)}) \]

therefore

\[ f(x) - f(x^+) \geq \frac{t}{2nR^2} (f(x) - f^*)^2 \]
let \( \Delta_k = f(x^{(k)}) - f^* \), then

\[
\Delta_k - \Delta_{k+1} \geq \frac{t}{2nR^2}\Delta_k^2
\]

consider their multiplicative inverses

\[
\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} = \frac{\Delta_k - \Delta_{k+1}}{\Delta_{k+1}\Delta_k} \geq \frac{\Delta_k - \Delta_{k+1}}{\Delta_k^2} \geq \frac{t}{2nR^2}
\]

therefore

\[
\frac{1}{\Delta_k} \geq \frac{1}{\Delta_0} + \frac{k}{2nL_{\max}R^2} \geq \frac{2t}{nR^2} + \frac{kt}{2nR^2}
\]

finally

\[
f(x^{(k)}) - f^* = \Delta_k \leq \frac{2nR^2}{(k + 4)t}
\]
Bounds on full gradient Lipschitz constant

**Lemma:** Suppose \( A \in \mathbb{R}^{N \times N} \) is positive semidefinite and has the partition \( A = [A_{ij}]_{n \times n} \), where \( A_{ij} \in \mathbb{R}^{N_i \times N_j} \) for \( i, j = 1, \ldots, n \), and

\[ A_{ii} \preceq L_i I_{N_i}, \quad i = 1, \ldots, n \]

then

\[ A \preceq \left( \sum_{i=1}^{n} L_i \right) I_N \]

**Proof:**

\[ x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i^T A_{ij} x_j \leq \left( \sum_{i=1}^{n} \sqrt{x_i^T A_{ii} x_i} \right)^2 \leq \left( \sum_{i=1}^{n} L_i \right) \sum_{i=1}^{n} \| x_i \|_2^2 \]

**Conclusion:** The full gradient Lipschitz constant \( L_f \leq \sum_{i=1}^{n} L_i \)
Computational complexity and justifications

(steepest) coordinate descent \( O\left(\frac{nMR^2}{k}\right) \)

full gradient method \( O\left(\frac{L_fR^2}{k}\right) \)

in general coordinate descent has worse complexity bound

- it can happen that \( M \geq O(L_f) \)
- choosing \( i(k) \) may rely on computing full gradient
- too expensive to do line search based on function values

nevertheless, there are justifications for huge-scale problems

- even computation of a function value can require substantial effort
- limits by computer memory, distributed storage, and human patience
Example

\[
\minimize_{x \in \mathbb{R}^n} \left\{ f(x) \overset{\text{def}}{=} \sum_{i=1}^n f_i(x_i) + \frac{1}{2} \|Ax - b\|_2^2 \right\}
\]

- \( f_i \) are convex differentiable univariate functions
- \( A = [a_1 \cdots a_n] \in \mathbb{R}^{m \times n} \), and assume \( a_i \) has \( p_i \) nonzero elements

computing either function value or full gradient costs \( O(\sum_{i=1}^n p_i) \) operations

**computing coordinate directional derivatives:** \( O(p_i) \) operations

\[
\nabla_i f(x) = \nabla f_i(x_i) + a_i^T r(x), \quad i = 1, \ldots, n
\]

\[
r(x) = Ax - b
\]

- given \( r(x) \), computing \( \nabla_i f(x) \) requires \( O(p_i) \) operations
- coordinate update \( x^+ = x + \alpha e_i \) results in efficient update of residue:
  \[
  r(x^+) = r(x) + \alpha a_i, \text{ which also cost } O(p_i) \text{ operations}
  \]
Outline

• theoretical justifications
• randomized coordinate descent method
• minimizing composite objectives
• accelerated coordinate descent method
Randomized coordinate descent

choose \( x^{(0)} \in \mathbb{R}^n \) and \( \alpha \in \mathbb{R} \), and iterate for \( k = 0, 1, 2, \ldots \)

1. choose \( i(k) \) with probability

\[
p_i^{(\alpha)} = \frac{L_i^\alpha}{\sum_{j=1}^{n} L_j^\alpha}, \quad i = 1, \ldots, n
\]

2. update

\[
x^{(k+1)} = x^{(k)} - \frac{1}{L_i} U_{i(k)} \nabla_i f(x^{(k)})
\]

special case: \( \alpha = 0 \) gives uniform distribution \( p_i^{(0)} = 1/n \) for \( i = 1, \ldots, n \)

assumptions

• \( \nabla f(x) \) is block-wise Lipschitz continuous

\[
\| \nabla_i f(x + U_i v_i) - \nabla_i f(x) \|_2 \leq L_i \| v_i \|_2, \quad i = 1, \ldots, n
\]  

(1)

• \( f \) has bounded sub-level set, and \( f^* \) is attained at some \( x^* \)
Solution guarantees

convergence in expectation

\[ \mathbb{E}[f(x^{(k)})] - f^* \leq \epsilon \]

high probability iteration complexity: number of iterations to reach

\[ \text{prob}(f(x^{(k)}) - f^* \leq \epsilon) \geq 1 - \rho \]

- confidence level \( 0 < \rho < 1 \)
- error tolerance \( \epsilon > 0 \)
Convergence analysis

block coordinate-wise Lipschitz continuity of $\nabla f(x)$ implies for $i = 1, \ldots, n$

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

coordinate update obtained by minimizing quadratic upper bound

$$x^+ = x + U_i \hat{v}_i$$

$$\hat{v}_i = \arg\min_{v_i} \left\{ f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2} \|v_i\|^2 \right\}$$

objective function is non-increasing:

$$f(x) - f(x^+) \geq \frac{1}{2L_i} \|\nabla_i f(x)\|^2_2$$
A pair of conjugate norms

for any \( \alpha \in \mathbb{R} \), define

\[
\| x \|_{\alpha} = \left( \sum_{i=1}^{n} L_i^{\alpha} \| x_i \|_2^2 \right)^{1/2}, \quad \| y \|_{\alpha}^* = \left( \sum_{i=1}^{n} L_i^{-\alpha} \| y_i \|_2^2 \right)^{1/2}
\]

let \( S_{\alpha} = \sum_{i=1}^{n} L_i^{\alpha} \) (note that \( S_0 = n \))

**lemma** (Nesterov): let \( f \) satisfy (1), then for any \( \alpha \in \mathbb{R} \),

\[
\| \nabla f(x) - \nabla f(y) \|_{1-\alpha}^* \leq S_{\alpha} \| x - y \|_{1-\alpha}, \quad \forall x, y \in \mathbb{R}^N
\]

therefore

\[
f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{S_{\alpha}}{2} \| x - y \|_{1-\alpha}^2, \quad \forall x, y \in \mathbb{R}^N
\]
Convergence in expectation

**Theorem** (Nesterov): for any \( k \geq 0 \),

\[
E f(x(k)) - f^* \leq \frac{2}{k + 4} S_\alpha R_{1-\alpha}^2(x^{(0)})
\]

where \( R_{1-\alpha}(x^{(0)}) = \max_y \left\{ \max_{x^* \in X^*} \|y - x^*\|_{1-\alpha} : f(y) \leq f(x^{(0)}) \right\} \)

**Proof:** define random variables \( \xi_k = \{i(0), \ldots, i(k)\} \),

\[
f(x^{(k)}) - E_{i(k)} f(x^{(k+1)}) = \sum_{i=1}^{n} p_i^{(\alpha)} (f(x^{(k)}) - f(x^{(k)} + U_i \hat{v}_i))
\]

\[
\geq \sum_{i=1}^{n} \frac{p_i^{(\alpha)}}{2L_i} \|\nabla_i f(x^{(k)})\|_2^2
\]

\[
= \frac{1}{2S_\alpha} (\|\nabla f(x)\|_1^{* - \alpha})^2
\]

Coordinate descent methods
\[ f(x^{(k)}) - f^* \leq \min_{x^* \in X^*} \langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle \leq \| \nabla f(x^{(k)}) \|^*_1 \| R_{1-\alpha}(x^{(0)}) \|^{1-\alpha} \]

therefore, with \( C = 2S\alpha R_{1-\alpha}(x^{(0)}) \),

\[ f(x^{(k)}) - \mathbb{E}_{i^{(k)}} f(x^{(k+1)}) \geq \frac{1}{C} (f(x^{(k)}) - f^*)^2 \]

taking expectation of both sides with respect to \( \xi_{k-1} = \{i(0), \ldots, i(k-1)\} \),

\[ \mathbb{E} f(x^{(k)}) - \mathbb{E} f(x^{(k+1)}) \geq \frac{1}{C} \mathbb{E}_{\xi_{k-1}} \left[ (f(x^{(k)}) - f^*)^2 \right] \geq \frac{1}{C} \left( \mathbb{E} f(x^{(k)}) - f^* \right)^2 \]

finally, following steps on page 12–6 to obtain desired result
Discussions

• $\alpha = 0$: $S_0 = n$ and

$$
\mathbb{E} f(x^{(k)}) - f^* \leq \frac{2n}{k + 4} R_1^2(x^{(0)}) \leq \frac{2n}{k + 4} \sum_{i=1}^{n} L_i \|x_i^{(0)} - x^*\|_2^2
$$

corresponding rate of full gradient method: $f(x^{(k)}) - f^* \leq \gamma R_1^2(x^{(0)})$, where $\gamma$ is big enough to ensure $\nabla^2 f(x) \preceq \gamma \text{diag}\{L_i I_{N_i}\}_{i=1}^n$

c**onclusion**: proportional to worst case rate of full gradient method

• $\alpha = 1$: $S_1 = \sum_{i=1}^{n} L_i$ and

$$
\mathbb{E} f(x^{(k)}) - f^* \leq \frac{2}{k + 4} \left( \sum_{i=1}^{n} L_i \right) R_0^2(x^{(0)})
$$

corresponding rate of full gradient method: $f(x^{(k)}) - f^* \leq \frac{L_f}{k} R_0^2(x^{(0)})$

c**onclusion**: same as worst case rate of full gradient method

but each iteration of randomized coordinate descent can be much cheaper
An interesting case

consider $\alpha = 1/2$, let $N_i = 1$ for $i = 1, \ldots, n$, and let

$$D_\infty(x^{(0)}) = \max_x \left\{ \max_{y \in X^*} \max_{1 \leq i \leq n} |x_i - y_i| : f(x) \leq f(x^{(0)}) \right\}$$

then $R_{1/2}^2(x^{(0)}) \leq S_{1/2} D^2_\infty(x^{(0)})$ and hence

$$E f(x^{(k)}) - f^* \leq \frac{2}{k + 4} \left( \sum_{i=1}^{n} L_i^{1/2} \right)^2 D^2_\infty(x^{(0)})$$

• worst-case dimension-independent complexity of minimizing convex functions over $n$-dimensional box is infinite (Nemirovski & Yudin 1983)
• $S_{1/2}$ can be bounded for very big or even infinite dimension problems

**Conclusion:** RCD can work in situations where full gradient methods have no theoretical justification
Convergence for strongly convex functions

**Theorem** (Nesterov): if $f$ is strongly convex with respect to the norm $\| \cdot \|_{1-\alpha}$ with convexity parameter $\sigma_{1-\alpha} > 0$, then

$$Ef(x^{(k)}) - f^* \leq \left(1 - \frac{\sigma_{1-\alpha}}{S_\alpha}\right)^k (f(x^{(0)}) - f^*)$$

**Proof:** combine consequence of strong convexity

$$f(x^{(k)}) - f^* \leq \frac{1}{\sigma_{1-\alpha}} \left(\|\nabla f(x)\|_{1-\alpha}^*\right)^2$$

with inequality on page 12–14 to obtain

$$f(x^{(k)}) - Ef_i(x^{(k+1)}) \geq \frac{1}{2S_\alpha} \left(\|\nabla f(x)\|_{1-\alpha}^*\right)^2 \geq \frac{\sigma_{1-\alpha}}{S_\alpha} (f(x^{(k)}) - f^*)$$

it remains to take expectations over $\xi_{k-1} = \{i(0), \ldots, i(k-1)\}$

Coordinate descent methods
High probability bounds

number of iterations to guarantee

\[ \text{prob}(f(x^{(k)}) - f^* \leq \epsilon) \geq 1 - \rho \]

where \(0 < \rho < 1\) is confidence level and \(\epsilon > 0\) is error tolerance

- for smooth convex functions

\[ O\left(\frac{n}{\epsilon} \left(1 + \log \frac{1}{\rho}\right)\right) \]

- for smooth strongly convex functions

\[ O\left(\frac{n}{\mu} \log\left(\frac{1}{\epsilon \rho}\right)\right) \]
Outline

- theoretical justifications
- randomized coordinate descent method
- minimizing composite objectives
- accelerated coordinate descent method
Minimizing composite objectives

\[
\begin{align*}
\text{minimize} \quad & \{ F(x) \triangleq f(x) + \Psi(x) \} \\
\text{subject to} \quad & x \in \mathbb{R}^N
\end{align*}
\]

assumptions

- \( f \) differentiable and \( \nabla f(x) \) block coordinate-wise Lipschitz continuous

\[ \| \nabla_i f(x + U_i v_i) - \nabla_i f(x) \|_2 \leq L_i \| v_i \|_2, \quad i = 1, \ldots, n \]

- \( \Psi \) is block separable:

\[ \Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i) \]

and each \( \Psi_i \) is convex and closed, and also simple
Coordinate update

use quadratic upper bound on smooth part:

\[
F(x + U_i v) = f(x + U_i v_i) + \Psi(x + U_i v_i)
\]

\[
\leq f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2} \|v_i\|_2 + \Psi_i(x_i + v_i) + \sum_{j \neq i} \Psi_j(x_j)
\]

define

\[
V_i(x, v_i) = f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2} \|v_i\|_2 + \Psi_i(x_i + v_i)
\]

coordinate descent takes the form

\[
x^{(k+1)} = x^{(k)} + U_i \Delta x_i
\]

where

\[
\Delta x_i = \arg\min_{v_i} V(x, v_i)
\]
Randomized coordinate descent for composite functions

choose $x^{(0)} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, and iterate for $k = 0, 1, 2, \ldots$

1. choose $i(k)$ with uniform probability $1/n$
2. compute $\Delta x_i = \arg\min_v V(x^{(k)}, v_i)$ and update
   $$x^{(k+1)} = x^{(k)} + U_i \Delta x_i$$

• similar convergence results as for the smooth case

• can only choose coordinate with uniform distribution?

(see references for details)
Outline

• theoretical justifications

• randomized coordinate descent method

• minimizing composite objectives

• accelerated coordinate descent method
Assumptions

restrict to unconstrained smooth minimization problem

\[
\minimize_{x \in \mathbb{R}^N} f(x)
\]

assumptions

• \( \nabla f(x) \) is block-wise Lipschitz continuous

\[
\| \nabla_i f(x + U_i v) - \nabla_i f(x) \|_2 \leq L_i \| v \|_2, \quad i = 1, \ldots, n
\]

• \( f \) has convexity parameter \( \mu \geq 0 \)

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \| y - x \|_2^2
\]
Algorithm: ARCD($x^0$)

Set $v^0 = x^0$, choose $\gamma_0 > 0$ arbitrarily, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu$$

and set $\gamma_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu$

2. Compute $y^k = \frac{1}{\frac{\alpha_k}{n} \gamma_k + \gamma_{k+1}} \left(\frac{\alpha_k}{n} \gamma_k v^k + \gamma_{k+1} x^k\right)$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \frac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)$$

4. Set $v^{k+1} = \frac{1}{\gamma_{k+1}} \left(\left(1 - \frac{\alpha_k}{n}\right)\gamma_k v^k + \frac{\alpha_k}{n}\mu y^k - \frac{\alpha_k}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)\right)$
Algorithm: ARCD(\(x^0\))

Set \(v^0 = x^0\), choose \(\alpha_{-1} \in (0, n]\), and repeat for \(k = 0, 1, 2,\ldots\)

1. Compute \(\alpha_k \in (0, n]\) from the equation

\[
\alpha_k^2 = (1 - \frac{\alpha_k}{n}) \alpha_{k-1}^2 + \frac{\alpha_k}{n} \mu,
\]

and set \(\theta_k = \frac{n\alpha_k - \mu}{n^2 - \mu}\), \(\beta_k = 1 - \frac{\mu}{n\alpha_k}\)

2. Compute \(y^k = \theta_k v^k + (1 - \theta_k) x^k\)

3. Choose \(i_k \in \{1, \ldots, n\}\) uniformly at random, and update

\[
x^{k+1} = y^k - \frac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)
\]

4. Set \(v^{k+1} = \beta_k v^k + (1 - \beta_k) y^k - \frac{1}{\alpha_k L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)\)
**Convergence analysis**

**Theorem:** Let $x^*$ be a solution of $\min_x f(x)$ and $f^*$ be the optimal value. If $\{x^k\}$ is generated by ARCD method, then for any $k \geq 0$

$$
\mathbb{E}[f(x^k)] - f^* \leq \lambda_k \left( f(x^0) - f^* + \frac{\gamma_0}{2} \|x^0 - x^*\|_2^2 \right),
$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1} \left( 1 - \frac{\alpha_i}{n} \right)$. In particular, if $\gamma_0 \geq \mu$, then

$$
\lambda_k \leq \min \left\{ \left( 1 - \frac{\sqrt{\mu}}{n} \right)^k, \left( \frac{n}{n + k \frac{\sqrt{\gamma_0}}{2}} \right)^2 \right\}.
$$

- when $n = 1$, recovers results for accelerated full gradient methods
- efficient implementation possible using change of variables

Coordinate descent methods
Randomized estimate sequence

definition: \( \{ (\phi_k(x), \lambda_k) \}_{k=0}^{\infty} \) is a randomized estimate sequence of \( f(x) \) if

- \( \lambda_k \to 0 \) (assume \( \lambda_k \) independent of \( \xi_k = \{i_0, \ldots, i_k\} \))
- \( E_{\xi_{k-1}}[\phi_k(x)] \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x), \ \forall x \in \mathbb{R}^N \)

lemma: if \( \{x^{(k)}\} \) satisfies \( E_{\xi_{k-1}}[f(x^k)] \leq \min_x E_{\xi_{k-1}}[\phi_k(x)] \), then

\[
E_{\xi_{k-1}}[f(x^k)] - f^* \leq \lambda_k (\phi_0(x^*) - f^*) \to 0
\]

proof:
\[
E_{\xi_{k-1}}[f(x^k)] \leq \min_x E_{\xi_{k-1}}[\phi_k(x)] \\
\leq \min_x \{ (1 - \lambda_k)f(x) + \lambda_k\phi_0(x) \} \\
\leq (1 - \lambda_k)f(x^*) + \lambda_k\phi_0(x^*) \\
= f^* + \lambda_k(\phi_0(x^*) - f^*)
\]
Construction of randomized estimate sequence

Lemma: if \[ \{\alpha_k\}_{k\geq 0} \] satisfies \[ \alpha_k \in (0, n) \] and \[ \sum_{k=0}^{\infty} \alpha_k = \infty \], then

\[ \lambda_{k+1} = \left( 1 - \frac{\alpha_k}{n} \right) \lambda_k, \quad \text{with} \quad \lambda_0 = 1 \]

\[ \phi_{k+1}(x) = \left( 1 - \frac{\alpha_k}{n} \right) \phi_k(x) + \frac{\alpha_k}{n} \left( f(y^k) + n\langle \nabla_i f(y^k), x_{ik} - y_{ik} \rangle + \frac{\mu}{2} \| x - y^k \|_L^2 \right) \]

is a pair of randomized estimate sequence

Proof: for \( k = 0 \), \( \mathbb{E}_{\xi_{k-1}}[\phi_0(x)] = \phi_0(x) = (1 - \lambda_0)f(x) + \lambda_0\phi_0(x) \); then

\[ \mathbb{E}_{\xi_k}[\phi_{k+1}(x)] = \mathbb{E}_{\xi_{k-1}} \left[ \mathbb{E}_{\xi_k}[\phi_{k+1}(x)] \right] \]

\[ = \mathbb{E}_{\xi_{k-1}} \left[ \left( 1 - \frac{\alpha_k}{n} \right) \phi_k(x) + \frac{\alpha_k}{n} \left( f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \| x - y^k \|_L^2 \right) \right] \]

\[ \leq \mathbb{E}_{\xi_{k-1}} \left[ \left( 1 - \frac{\alpha_k}{n} \right) \phi_k(x) + \frac{\alpha_k}{n} f(x) \right] \]

\[ \leq \left( 1 - \frac{\alpha_k}{n} \right) \left[ (1 - \lambda_k)f(x) + \lambda_k\phi_0(x) \right] + \frac{\alpha_k}{n} f(x) \quad \ldots \]
Derivation of APCD

• let \( \phi_0(x) = \phi^* + \gamma_0 \|x - v^0\|^2_L \), then for all \( k \geq 0 \),

\[
\phi_k(x) = \phi^*_k + \frac{\gamma_k}{2} \|x - v^k\|^2_L
\]

can derive expressions for \( \phi^*_k, \gamma_k \) and \( v^k \) explicitly

• follow the same steps as in deriving accelerated full gradient method

• actually use a strong condition

\[
\mathbb{E}_{\xi_{k-1}} f(x^k) \leq \mathbb{E}_{\xi_{k-1}} \left[ \min_x \phi_k(x) \right]
\]

which implies

\[
\mathbb{E}_{\xi_{k-1}} f(x^k) \leq \min_x \mathbb{E}_{\xi_{k-1}} \left[ \phi_k(x) \right]
\]
References


