Incremental Methods for Additive Convex Cost Minimization: Deterministic vs Randomized Variants

Mert Gurbuzbalaban (Rutgers)

joint work with

A. Ozdaglar (MIT), **P. Parrilo** (MIT), **D. Vanli** (MIT)

DIMACS Workshop, August 2017



Additive Cost Problems

- We consider optimization problems with an objective function given by the sum of a large number of component functions:

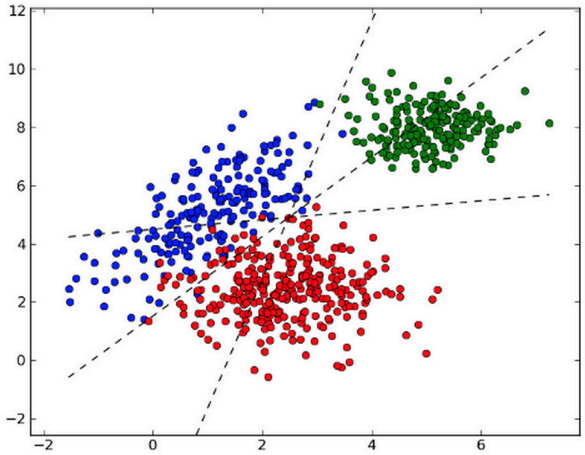
$$\begin{aligned} \min_{x} \quad & f(x) = \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \end{aligned}$$

  where $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ are convex functions.

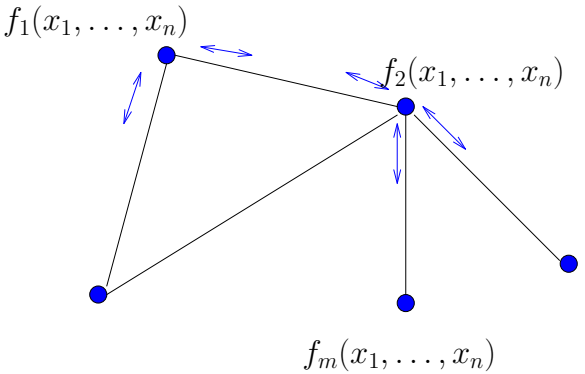
- These arise in several important contexts.

Examples of Additive Cost Problems

- Empirical Risk Minimization:
  - Data $\{(x_i, y_i)\}_{i=1}^m$: $x_i \in \mathbb{R}^n$ is a feature vector, $y_i \in \mathbb{R}$ is target output.
  - $\min_{\theta \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m L(y_i, x_i, \theta) + \text{pen}(\theta)$.
  - Examples: LASSO, support vector machine, logistic regression, classification...



- Minimization of an Expected Value (Stochastic Programming):
  - $\min_{x \in X} E[F(x, w)]$ (w: random variable taking large finite number of values).
- Distributed Optimization in Networks:
  - $f_i(x)$: local objective function of node i (privately known by node i).



Incremental Methods

- We focus on problems where the number of component functions m is large, so a full (sub)gradient step, $\nabla f(x) = \sum_{i=1}^{m} \nabla f_i(x)$, is very costly.
- Motivates using incremental algorithms which process component functions sequentially.
  - Reasonable progress with cheaper "incremental" steps.
- Also well-suited for problems where:
  - $f_i(x)$: distributed and locally known by agents.
  - $f_i(x)$: known sequentially over time in an online manner.
- Incremental Gradient: Each (outer) iteration k consists of a cycle with m subiterations: For $k \geq 1$,

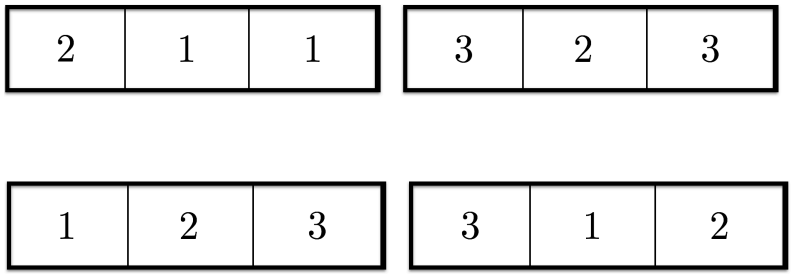
$$x_{i+1}^k = x_i^k - \alpha_k \nabla f_i(x_i^k), \quad \text{for} \quad i = 1, 2, \ldots, m,$$

where α_k is a stepsize.

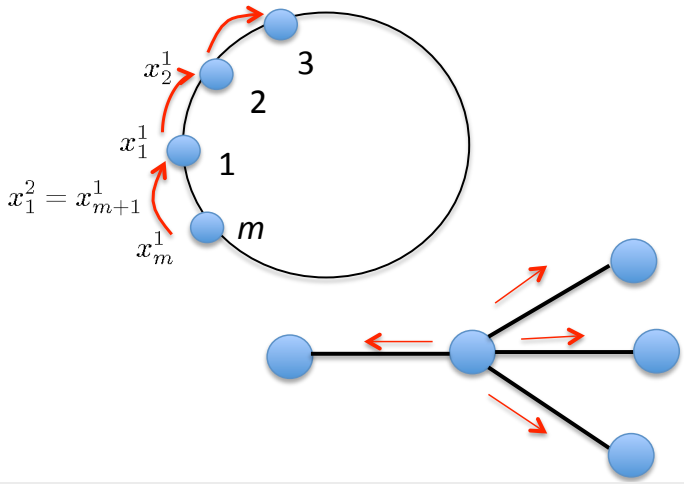
Order for Processing Component Functions

- Deterministic Orders:
  - Cyclic order: Incremental Gradient
  - Fixed arbitrary order in each cycle

- Random Orders:
  - Sample with replacement: Stochastic Gradient Descent (SGD)
  - Sample without replacement: Random Reshuffling (RR)



- Network-imposed Orders:
  - Deterministic with network structure.
  - Random (next component function sampled from neighborhood): Markov Randomized Incremental Methods.



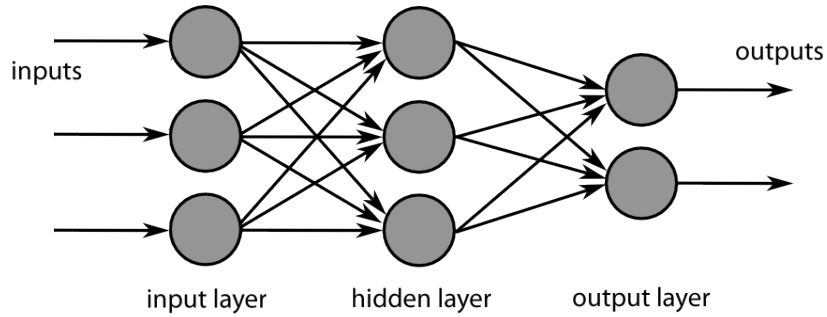
This Talk

- We study Incremental Gradient (IG) method for deterministic orders.
  - For smooth/strongly convex functions, we show $\mathcal{O}(1/k)$ rate in distances [$\mathcal{O}(1/k^2)$ rate in function values].
  - Improves on the existing $\mathcal{O}(1/\sqrt{k})$ result (for non smooth functions).
  - Achieving this rate with IG involves knowing strong convexity constant.
- We then focus on random orders, in particular Random Reshuffling (RR).
  - Numerically observed to outperform SGD, yet no analytical results.
  - We show $\Theta(1/k^{2s})$ rate, $s \in (1/2, 1)$, with probability one in function values.
  - Improves on the existing $\Omega(1/k)$ minmax rate of SGD.
  - Achieving this rate involves a stepsize $\alpha_k = 1/k^s$ and properly averaging the iterates.
- As a special case of IG, we study coordinate descent methods. We provide linear rate results and problem classes for which any cyclic order is faster than randomized order both asymptotically and non-asymptotically in the worst-case. We also characterize the best deterministic order.

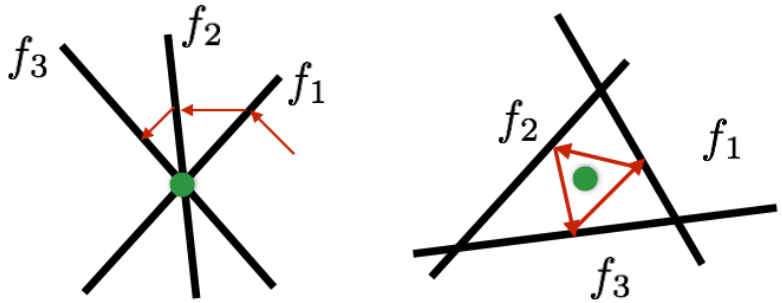
Incremental (Sub)Gradient method

Prominent algorithm that appears in many contexts:

- Backpropagation algorithm for training neural networks.



- Kaczmarz method for solving linear systems of equations $a_i^T x = b_i$.



$$f_i(x) = (a_i^T x - b_i)^2$$

Literature: Incremental (Sub)gradient Optimization

Deterministic order: Convergence analysis under various conditions

- Textbooks by Bertsekas, Polyak, Shor,...
- Differentiable problems: [Luo 91], [Luo and Tseng 94], [Mangasarian and Solodov 94], [Bertsekas 97], [Solodov 98], [Tseng 98],...
- Non-differentiable problems: [Nedic, Bertsekas 00], [Kiwiel 2004], ...
  - Best rate known $\text{dist}_k \leq \mathcal{O}(1/\sqrt{k})$ under strong-convexity-type cond.

Question: Can we achieve better rates when functions f_i are smooth?

Incremental Gradient with Smoothness

Assumptions:

1. (Strong convexity+differentiability) Each f_i is convex and C^2 on $\mathbb{R}^n$. The sum f is c-strongly convex, i.e.

$$f(x) - \frac{c}{2}\|x\|^2 \quad \text{is convex.}$$

2. (Lipschitz gradients) There exists a constant $L_i > 0$ such that

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x - y\|, \quad \text{for all} \quad x, y, \quad i = 1, 2, \ldots, m.$$

   Then, f has Lipschitz gradients with constant at most $L = \sum_i L_i$.

3. (Subgradient boundedness)

$$\|g\| \leq G, \quad \forall g \in \partial f_i(x_i^k), \quad i = 1, 2, \ldots, m, \quad k = 1, 2, \ldots.$$

Convergence Rate of IG with Smoothness

Theorem (Gurbuzbalaban, Ozdaglar, Parrilo 15)

Suppose Assumptions 1, 2 and 3 hold. Consider the IG method with stepsize $\alpha_k = R/k$. *If* $R > 1/c$, *then*

$$dist_k \leq \left(\frac{LmGR^2}{Rc-1}\right)\frac{1}{k} + o(1/k).$$

- This rate result highly dependent on the choice of stepsize, i.e., knowledge of strong convexity constant c.
  - Similar problems with $1/k$-decay step sizes widely noted in stochastic approximation and stochastic gradient descent literatures [Chung 53], [Frees and Ruppert 87], [Nemirovsky, Juditsky, Lan, and Shapiro 09], [Bach and Moulines 11], [Bach 13].

Convergence Rate of IG with Smoothness

Example

Let $f_i(x) = x^2/20$ for $i = 1, 2$, $x \in \mathbb{R}$. Then, we have $m = 2$, $c = 1/5$ and $x^* = 0$. Take $R = 1$ which corresponds the stepsize $1/k$. The IG iterations are

$$x_1^{k+1} = \left(1 - \frac{1}{10k}\right)^2 x_1^k.$$

If $x_1 = 1$, a simple analysis shows $x_1^k = \text{dist}_k > \Omega(\frac{1}{k^{1/5}})$.

- The stepsize $\alpha_k = \Theta(1/k^s)$, $s \in (0, 1)$, does not require adaptation to the strong convexity constant, providing robust rate guarantees.

Theorem (Gurbuzbalaban, Ozdaglar, Parrilo 15)

Suppose Assumptions 1, 2 and 3 hold. Consider the IG method with stepsize $\alpha_k = R/k^s$, $s \in (0, 1)$, *with* $R > 0$. *Then*

$$dist_k \leq \left(\frac{LmGR}{c}\right)\frac{1}{k^s} + o(1/k^s).$$

Quadratics: Order-Dependent Upper Bounds

- Consider the IG method with arbitrary deterministic order σ (a fixed permutation of $\{1, 2, \ldots, m\}$), and with stepsize $\alpha_k = R/k^s$, $s \in (0, 1)$.

Theorem (Gurbuzbalaban, Ozdaglar, Parrilo 2015)

For each i, let $f_i : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions of the form

$$f_i(x) = \frac{1}{2} x_i^T P_i x - q_i^T x + r_i,$$

where P_i is a symmetric square matrix, q_i is a column vector and r_i is a scalar. Suppose f is strongly convex with constant c. Then,

$$dist_k \leq \frac{R M_\sigma}{c} \frac{1}{k^s} + o(1/k^s), \quad \text{where} \quad M_\sigma = \left\| \sum_{1 \leq i < j \leq m} P_{\sigma(j)} \nabla f_{\sigma(i)}(x^*) \right\|.$$

- Note that $M_\sigma \leq \sum_{j=1}^m j L_{\sigma(j)} G \leq LmG$.
  - Suggests processing functions with higher Lipschitz constants first.

Random Orders: SGD vs RR

- Much empirical evidence showing RR outperforms SGD, no analytical results.

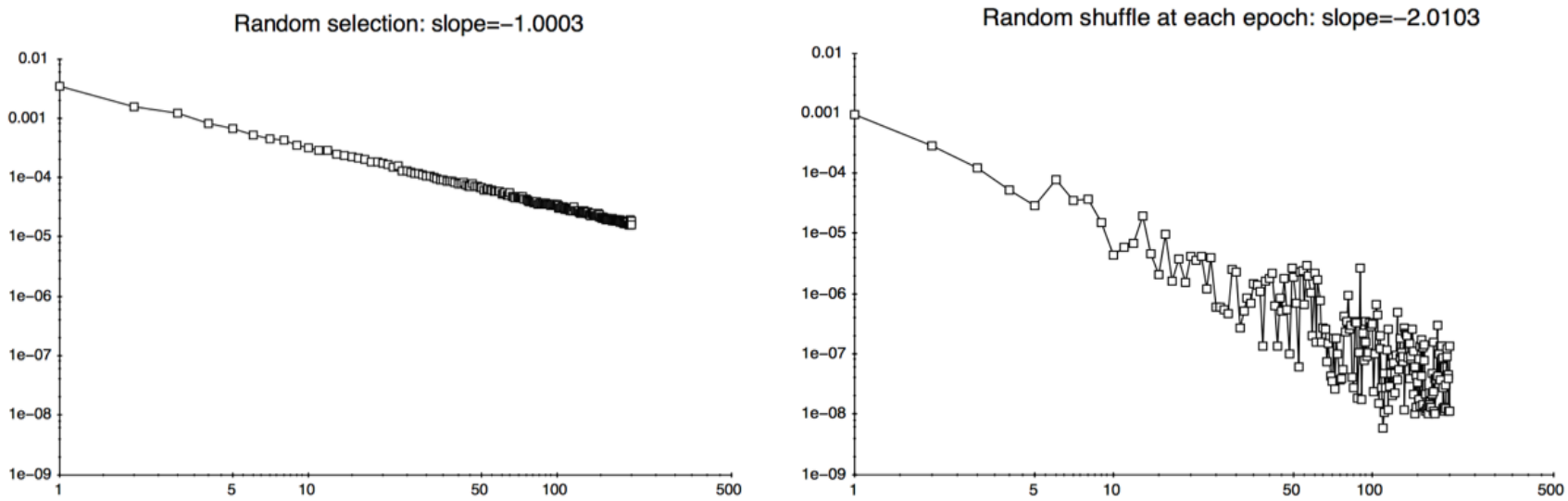


Figure: The classification of RCV1 documents belonging to class CCAT. Left: SGD achieves its $\Omega(1/k)$ rate, Right: Random Reshuffling rate of $\sim 1/k^2$ [Bottou 09].

- Long-standing open problem: Characterization of convergence rate of RR [Bertsekas 99], [Bottou 09], [Recht Re 2012, 2013].
- Analysis hard because of dependencies of gradient errors in and across cycles.

SGD: Revived Interest

- Vast literature going back to [Robbins, Monro 51], [Kiefer, Wolfovitz 52].
- Popular in machine learning applications due to its scalability and robustness.
- Active area of research: More recent work on achievable rates, more robust variants and second-order versions:

  [Ruppert 88], [Polyak 90], [Polyak, Juditsky 92], [Bottou, LeCun 05], [Nemirovski Juditsky, Lan and Shapiro 09], [Hazan, Kale 11], [Rakhlin, Shamir, Sridharan 12], [Bach and Moulines 11], [Byrd, Hansen, Nocedal, Singer 14], [Hardt, Recht, Singer 15]....

Convergence Rate of SGD

- For strongly convex functions, SGD has $\Omega(1/k)$ min-max lower bounds for stochastic convex optimization [Nemirovski, Yudin 83], [Agarwal et al. 12].
- Polyak-Ruppert averaging is one way of achieving this lower bound.
  - Choose larger stepsize $\alpha_k = R/k^s$ with $s \in (1/2, 1)$.
  - Take time average of the iterates

$$\bar{x}_k = \frac{x_1 + x_2 + \cdots + x_k}{k}$$

- **Averaged Stochastic Gradient Descent:**

Theorem (Polyak, Juditsky 92)

$$k^{1/2}(\bar{x}_k - x^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$$

$\Longrightarrow \sim 1/k$ *rate for function values.*

Convergence Rate of SGD and RR

Under Assumptions $1, 2$ + some technical conditions, we have:

- **Averaged Stochastic Gradient Descent:**

Theorem (Polyak, Juditsky 92)

$$k^{1/2}(\bar{x}_k - x^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$$

$\Longrightarrow \sim 1/k$ *rate for function values.*

- **Random Reshuffling (RR):**

Theorem (Gurbuzbalaban, Ozdaglar, Parrilo 15 (simplified))

$$k^s(\bar{x}_k - x^*) \to \nabla^2 f(x^*)^{-1}\theta_* \quad \textit{with probability one}$$

for a fixed vector $\theta_* = -\frac{1}{2}\sum_{i=1}^m \nabla^2 f_i(x^*)\nabla f_i(x^*)$ *and* $s \in (1/2, 1)$.

$\Longrightarrow$ $\sim 1/k^{2s}$ *faster rate for function values. Also,* $\|\theta_*\| \leq LG$ *(no additional* m*).*

Ilustration on a simple example

- Two quadratics: $f_1(x) = \frac{1}{2}(x+1)^2$, $f_2(x) = \frac{1}{2}(x-1)^2$. Here, $\theta^* = 0$.

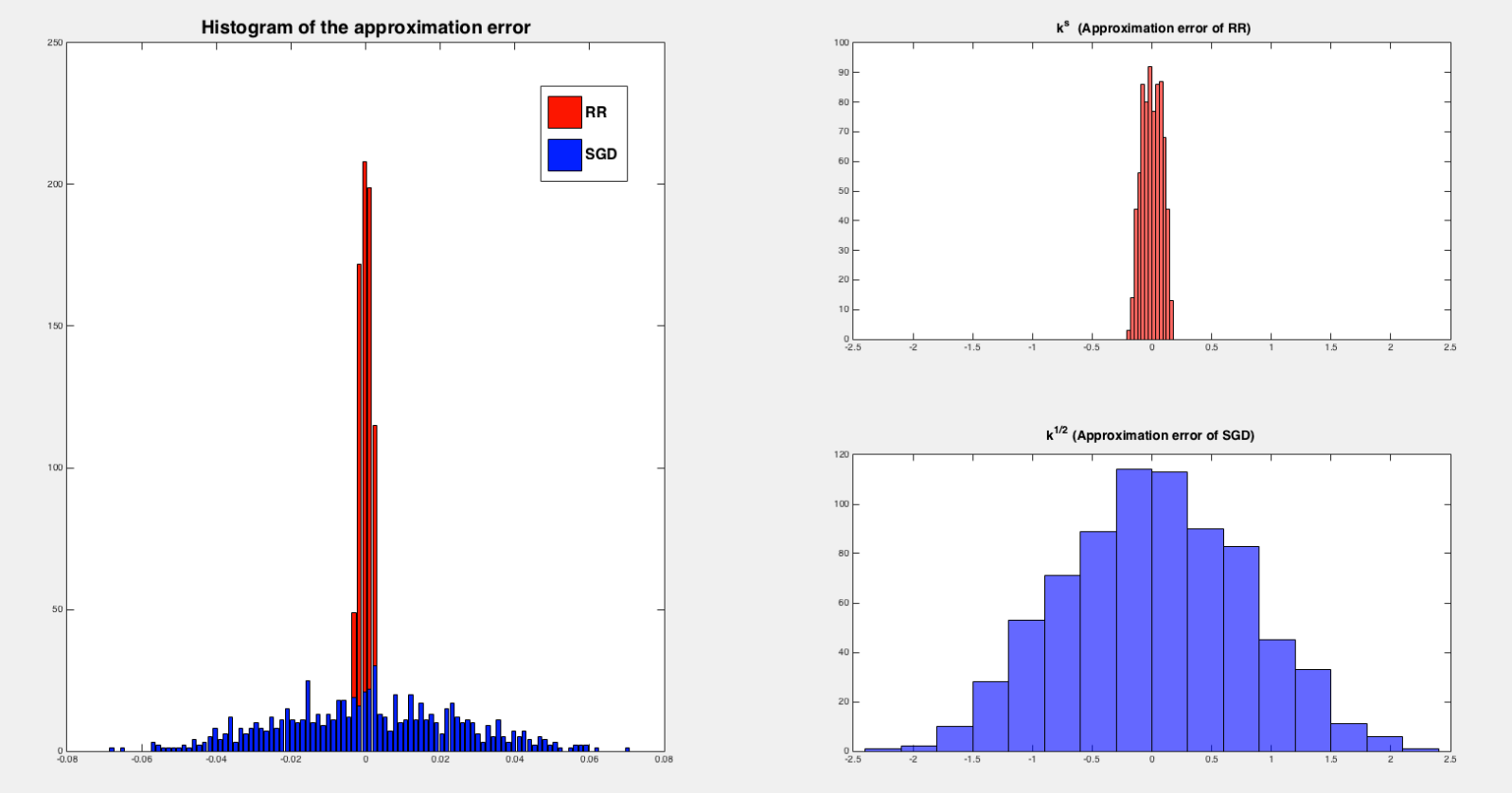


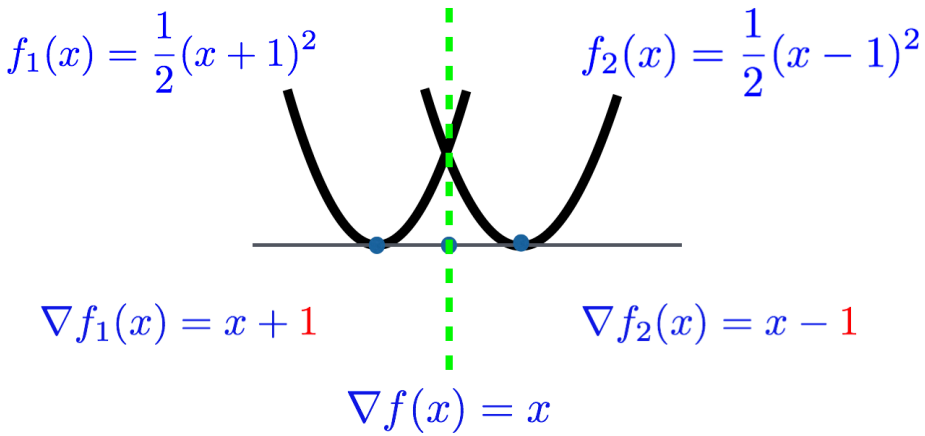
Figure: Left: Histograms of the approximation error $\Delta_k = \bar{x}_k - x^*$ for SGD and RR. Right, top: Histogram of $k^s \Delta_k \to 0$ for RR as $\theta^* = 0$. Right, bottom: Histogram of $k^{1/2}\Delta_k$ for SGD which is asymptotically normal.

Intuition: Bias-Variance Trade-Off

- **SGD:** samples index i_k uniformly and independently at iteration k.

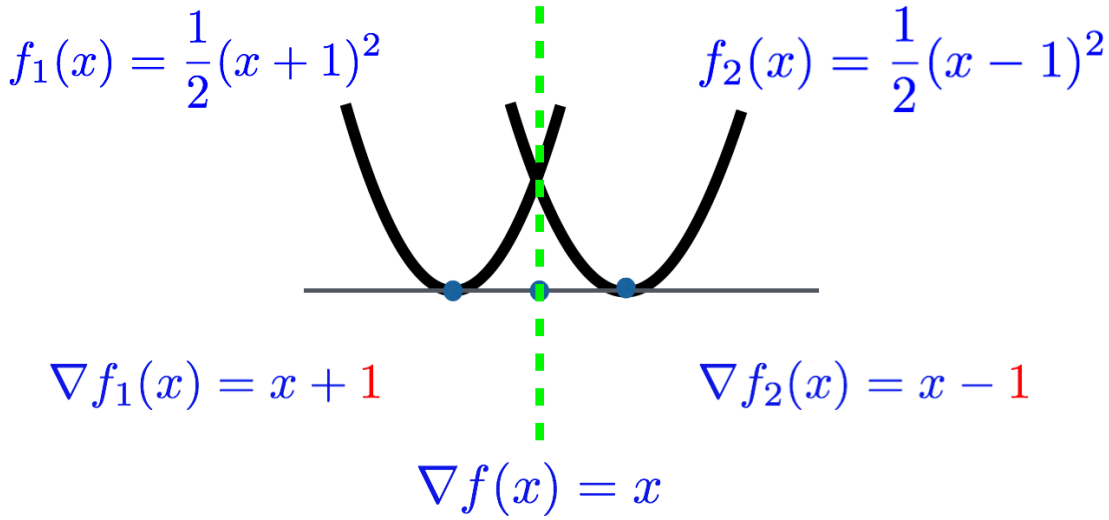
$$x^{k+1} = x^k - \alpha_k \nabla f_{i_k}(x^k) = x^k - \alpha_k(\nabla f(x^k) + E^k)$$

where E^k is the iteration gradient error.



- **SGD:** $E^k = \pm 1$ with prob $1/2$. $\mathbb{E}(E^k) = 0, \quad \text{var}(E^k) = 1.$
- The error sequence E^k is a martingale difference sequence.

Intuition: Bias-Variance Trade-Off

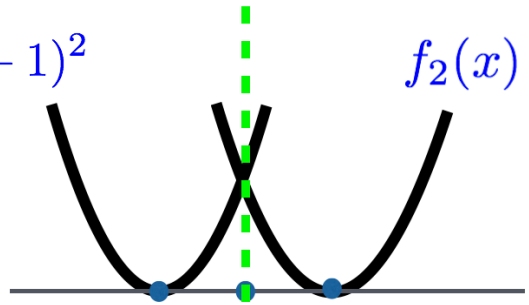


$$x_1^{k+1} = x_1^k - \alpha_k(\nabla f_1(x_1^k) + \nabla f_2(x_1^k) + e^k)$$

$$e^k = \begin{cases} \nabla f_2(x_2^k) - \nabla f_2(x_1^k) & \text{if} \quad \sigma_k = \{1,2\} \\ \nabla f_1(x_2^k) - \nabla f_1(x_1^k) & \text{if} \quad \sigma_k = \{2,1\} \end{cases}$$

- By gradient Lipschitzness: $e^k = O(\alpha_k), \quad \mathbb{E}(e^k) \neq 0, \quad \text{var}(e^k) = O(\alpha_k^2)$
- **RR error has reduced variance** but the error sequence e^k is not a martingale difference sequence due to correlations among the inner iterates.

Intuition: Bias-Variance Trade-Off



$$x_1^{k+1} = x_1^k - \alpha_k(\nabla f_1(x_1^k) + \nabla f_2(x_1^k) + e^k)$$

$$e^k = \begin{cases} \nabla f_2(x_2^k) - \nabla f_2(x_1^k) & \text{if} \quad \sigma_k = \{1,2\} \\ \nabla f_1(x_2^k) - \nabla f_1(x_1^k) & \text{if} \quad \sigma_k = \{2,1\} \end{cases} \implies e^k = \alpha_k v_k - \underbrace{\alpha_k(x_1^k - x^*)}_{\mathcal{O}(\alpha_k^2) \text{ by cyclic analysis}}$$

where $v_k = v(\sigma_k)$ is a sequence independent over cycles.

- By gradient Lipschitzness: $e^k = O(\alpha_k), \quad \mathbb{E}(e^k) \neq 0, \quad \text{var}(e^k) = O(\alpha_k^2)$
- **RR error has reduced variance** but the error sequence e^k is not a martingale difference sequence due to correlations among the inner iterates.

Proof Sketch (specialize to quadratics):

- Evolution of outer RR iterates is given by

$$\frac{x_1^k - x_1^{k+1}}{\alpha_k} = \nabla f(x_1^k) + e^k,$$

where e^k is the cycle gradient error.

- Averaging both sides and using $\nabla f(x_1^j) = H_*(x_1^j - x^*)$ (with $H_* = \nabla^2 f(x^*)$),

$$I_k := \frac{\sum_{j=0}^{k-1}(x_1^j - x_1^{j+1})\alpha_j^{-1}}{k} = \frac{\sum_{j=0}^{k-1} H_*(x_1^j - x^*) + e^j}{k}.$$

- Equivalently,

$$\bar{x}_k - x^* = -H_*^{-1} \underbrace{\bar{\alpha}_k}_{\mathcal{O}\left(\frac{1}{k^s}\right)} \underbrace{\frac{\sum_j e^j}{\sum_j \alpha_j}}_{\to \theta^* \ \ a.s.} + H_*^{-1} \underbrace{I_k}_{\mathcal{O}\left(\frac{\log k}{k}\right)},$$

where $\bar{\alpha}_k = \sum_j \alpha_j / k$ is the averaged stepsize.

Proof Sketch (specialize to quadratics):

- $\mathcal{O}\left(\frac{\log k}{k}\right)$: follows from deterministic IG results and "lots of algebra".
- $\frac{\sum_j e^j}{\sum_j \alpha_j} \to \theta^*$ a.s.: follows from decomposing the cycle gradient error:

$$e^k = \alpha_k v_k + \mathcal{O}(\alpha_k^2),$$

where v_k is a sequence independent over cycles with

$$E[v_k] := \theta_* = \frac{1}{2}\sum_{i=1}^{m} \nabla^2 f_i(x^*)\nabla f_i(x^*).$$

- By strong law of large numbers, we have $\frac{\sum_j \frac{e^j}{\alpha_j}}{k} \to E[v_k]$ a.s., implying almost sure convergence of the weighted version $\frac{\sum_j e^j}{\sum_j \alpha_j}$.

Accelerating RR Further: Bias Removal

- Bottleneck term:

$$\text{Deterministic bias } (k) := \bar{\alpha}_k H_*^{-1}\theta_*, \quad \theta_* = -\frac{1}{2}\sum_{i=1}^{m}\nabla^2 f_i(x^*)\nabla f_i(x^*).$$

- Estimate bias in last cycle and subtract to get $1/k^2$ rate in function values!

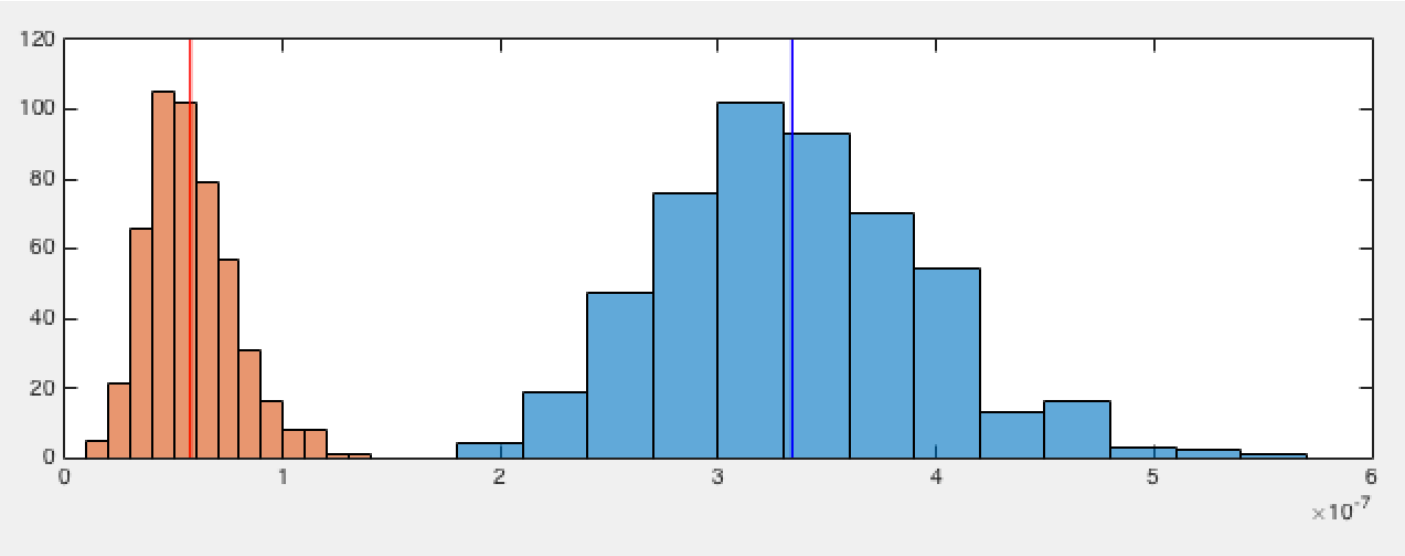


Figure: Histograms of the suboptimality of the function values for a fixed number of cycles. In Orange: Accelerated RR, In Blue: RR.

Special Case of IG: Coordinate Descent

- For $f : \mathbb{R}^n \to \mathbb{R}$ is convex & smooth, we consider unconstrained problems:

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

- CD algorithm: At each iteration k, select an index i_k and approximately minimize the objective in the i_k-th coordinate:

$$\begin{aligned} x^{k+1} &= x^k - \underbrace{\eta_k}_{\text{stepsize}} [\nabla f(x^k)]_{i_k} e_{i_k}. \\ [\nabla f(x)]_{i_k} &= i_k\text{-th component of the gradient } \nabla f(x) =: \nabla f_{i_k}(x) \\ e_{i_k} &= [0, 0, \ldots, 1, 0, \ldots 0]^T = \text{the } i_k\text{-th coordinate vector} \end{aligned}$$

- CD methods have a long history in optimization, their convergence properties have been studied extensively in late 70's to, 90's: [Bertsekas Tsitsiklis 89], [Bertsekas 99], [Tseng Luo 92], [Grippo Scandrione, 99], [Auslender 76].
- Resurgence of recent interest because of their applicability in machine learning as well as large scale data analysis and superior empirical performance.

Recent Work

- Choice of order i_k:
  - Deterministic Order: Cyclic Coordinate Descent (CCD)
    - [Beck, Tetruashvili 13], [Sun, Hong 15]: Global rate estimates, which suggests CCD is $O(n^2)$ times slower than RCD for strongly convex f.
    - Puzzling in view of the empirical faster performance of CCD over RCD for various problems.
    - [Sun, Ye 16]: Provided a quadratic problem for which the $O(n^2)$ gap in [Beck, Tetruashvili 13] is achieved.
  - Random Orders: Random CD (RCD), Randomly Permuted CD (RPCD)
    - [Nesterov 12]: Provided the first global non-asymptotic convergence rates of RCD for convex and smooth problems.
    - [Lee, Wright 16]: Tight analysis for RPCD on the quadratic example of [Sun, Ye 16].
- These results suggest that CCD is slower than RCD wrt scaling in n..
- Active research area including [Richtarik,Takac 11], [Scutari et al 14], [Wright 15], [Saha, Tewari 10], [Wang Lin], [Nesterov, Stich, 17], [Liu, Wright 16], [Lin, Lu, Xiao 14], [Hong et al 13], [Nutini et al. 15], [Necoara et al 11],...

Setup

- We focus on convex quadratic problems

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax \quad \text{where} \quad A \in \mathbb{R}^{n\times n}. \tag{1}$$

- Assumption 1.
  - (i) A is invertible, i.e. $\mu := \lambda_{\min}(A) > 0$.
  - (ii) The diagonals of A are all normalized to one[1].

$$A_{i,i} = 1, \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{2}$$

- By (i), the problem (1) has unique solution at $x^* = 0$.
- Let C and R be the iteration matrices of CCD and RCD.
- We consider two problem classes: *i*) A is an M-matrix, i.e., the off-diagonal entries of A are nonnegative (ex: solving Laplacian-like systems), *ii*) A is a 2-cyclic matrix.

[1] This is not restrictive as we could always put A into this form by scaling x easily.

CD Iterations: Close-up

- CCD iterations:
  - Rewrite $A = I - L - L^T$, $-L$ is the strictly lower diagonal part of A.
  - With standard cyclic rule $1, \ldots, n$ (i.e., $i_k = k \pmod n + 1$):

$$x_{\text{CCD}}^{(\ell+1)n} = C\, x_{\text{CCD}}^{\ell n}, \quad \text{where} \quad C = (D - L)^{-1} L^T. \tag{3}$$

  - Equivalent to one iteration of the Gauss-Seidel method for $Ax = 0$.
- RCD iterations:
  - i_k is random (sampled with-replacement).
  - The iterates evolve in expectation as

$$\mathbb{E} x_{\text{RCD}}^{(\ell+1)n} = R\, \mathbb{E} x_{\text{RCD}}^{\ell n} \quad \text{with} \quad R := \left(I - \frac{1}{n} A\right)^n. \tag{4}$$

Asymptotic Rate of Convergence - I

- We use the notion of the worst-case asymptotic convergence rate that has been studied extensively in the literature for iterative algorithms [Ortega Rheinboldt 70], [Varga 09], [Bertsekas Tsitsiklis 89].
- The reduction in distance to optimality at the worst-case for CCD:

$$\sup_{x^0} \frac{\left|\left|x_{\text{CCD}}^{\ell n} - x^*\right|\right|}{\left|\left|x_{\text{CCD}}^{0} - x^*\right|\right|} = \|C^\ell\|, \quad \|C^\ell\|^{1/\ell} \to \rho(C) \quad \text{as} \quad \ell \to \infty.$$

  where $\rho(\cdot)$ is the spectral radius.
- The worst-case asymptotic convergence rate is then

$$\text{Rate(CCD)} := \lim_{\ell \to \infty} \sup_{x_{\text{CCD}}^0 \in \mathbb{R}^n} -\frac{1}{\ell} \log \left(\frac{\left|\left|x_{\text{CCD}}^{\ell n} - x^*\right|\right|}{\left|\left|x_{\text{CCD}}^{0} - x^*\right|\right|} \right) = -\log\left(\rho(C)\right).$$

Asymptotic Rate of Convergence - II

- For RCD, analogously we define

$$\text{Rate(RCD)} := \lim_{\ell \to \infty} \sup_{x^0_{\text{RCD}} \in \mathbb{R}^n} -\frac{1}{\ell} \log \left(\frac{\left|\left| \mathbb{E}(x^{\ell n}_{\text{RCD}}) - x^* \right|\right|}{||x^0_{\text{RCD}} - x^*||} \right) = -\log\left(\rho(R)\right).$$

- The convergence of the expected distance to optimal solution $\left|\left| \mathbb{E}(x^{\ell n}_{\text{RCD}}) - x^* \right|\right|$ has been studied in the literature [Sun, Ye 16].
- Our results generalizes to other notions of convergence such as the convergence of $\mathbb{E}\left|\left| x^{\ell n}_{\text{RCD}} - x^* \right|\right|^2$.

- **Question:** When does CCD converge faster than RCD asymptotically, i.e. when is $\rho(C) < \rho(R)$?

A Motivating Example

- Consider the 4×4 symmetric matrix satisfying Assumption 1 with $\mu = 1/2$:

$$A = \begin{bmatrix} 1 & 0 & -1/4 & -1/4 \\ 0 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & 0 \\ -1/4 & -1/4 & 0 & 1 \end{bmatrix}. \tag{5}$$

- Then, CCD matrix has an explicit form $C = \begin{bmatrix} 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 1/8 & 1/8 \\ 0 & 0 & 1/8 & 1/8 \end{bmatrix}$.
- We check: $\rho(C) = 1/4$ and $\rho(R) = \rho\left(\left(I - \frac{1}{4}A\right)^4\right) = \left(1 - \frac{\mu}{4}\right)^4 \geq 1 - \mu = \frac{1}{2}$.
- Therefore,

$$\frac{\mathrm{Rate(CCD)}}{\mathrm{Rate(RCD)}} = \frac{-\log(\rho(C))}{-\log(\rho(R))} \geq \frac{-\log(1/2)}{-\log(1/4)} = 2$$

- **Question:** Is there a more general class of such examples?

Convergence Rate of RCD

Lemma

Suppose Assumption 1 holds. Then, the RCD algorithm satisfies

$$\rho(R) = \left(1 - \frac{\mu}{n}\right)^n \geq 1 - \mu$$

Proof:

- By Assumption 1, $\mu > 0$ and $\mathrm{tr}(A) = n$, which implies all eigenvalues of the matrix A/n are in the interval $(0, 1)$.
- Hence,

$$\rho(R) = \lambda_{\max}\left(\left(I - \frac{1}{n}A\right)^n\right) = \left(1 - \frac{1}{n}\lambda_{\min}(A)\right)^n = \left(1 - \frac{\mu}{n}\right)^n.$$

M-Matrices

Definition (M-matrix)

A real matrix A with $A_{i,j} \leq 0$ for all $i \neq j$ is an M-matrix if A is nonsingular and $A^{-1} \geq 0$.

- M-matrices arise in many contexts in optimization and iterative algorithms.
  - **Ex:** minimization of quadratic forms of graph Laplacians for spectral partitioning and semisupervised learning.

Definition (Irreducibility)

A matrix A is irreducible if it is not similar via a permutation to a block upper triangular matrix (that has more than one block of positive size).

- Irreducibility: key condition for Perron-Frobenius theory.

Spectral Radius of CCD Iteration Matrix for M-Matrices

Theorem

Suppose Assumption 1 holds and A is an irreducible M-matrix. Then, the iteration matrix of the CCD algorithm satisfies the following inequality

$$(1-\mu)^2 \leq \rho(C) \leq \frac{1-\mu}{1+\mu}, \tag{6}$$

where the inequality on the left holds with equality if and only if A is a consistently ordered matrix.

Definition (Consistent Ordering (Simplified form))

If the eigenvalues of $B_\alpha = \alpha L + \frac{1}{\alpha} L^T$ are independent of α, then A is said to be consistently ordered.

- As $\rho(R) \geq 1-\mu$ by Lemma 1, this theorem implies $\rho(C) < \rho(R)$.
- In order to prove the lower bound of this theorem, we use a modified version of a key result from [Varga 2009, Lemma 4.12].

Proof Sketch of Lower Bound in Inequality (6):

- Similar to the 4×4 example, one can show $C = (I - L)^{-1} L^T \geq 0$.
- By the Perron-Frobenius Theorem, $\lambda = \rho(C)$ and $\exists z \geq 0$

$$Cz = \lambda z \iff (\lambda L + L^T) z = \lambda z \iff \rho(\lambda L + L^T) = \lambda$$

- Suffices to solve the equation $\rho(\lambda L + L^T) = \lambda = \sqrt{\lambda}\, \rho\left(B_{\sqrt{\lambda}}\right)$.

Lemma (Varga 2009, Lemma 4.12)

Consider $B_\alpha = \alpha L + \frac{1}{\alpha} L^T$ *for* $\alpha \in (0, 1]$.

1. *If A is consistently ordered, by definition $\rho(B_{\sqrt{\lambda}})$ is a constant.*
2. *Else, $\rho(B_{\sqrt{\lambda}})$ is strictly decreasing on $\lambda \in (0, 1]$.*

Proof idea:

- Using the Perron-Frobenius Theorem, $\rho(B_\alpha) = \lim_{t \to \infty} [\mathrm{tr}(B_\alpha^t)]^{1/t}$.
- Compute the diagonals $[B_\alpha^t)]_{i,i}$ as a sum of all possible walks from i to itself in t steps.

Proof of Varga's Lemma

- As $B_\alpha \geq 0$ and B_α is irreducible, the largest eigenvalue of B_α has a multiplicity of 1. Therefore,

$$\rho(B_\alpha) = \lim_{t\to\infty} [\mathrm{tr}(B_\alpha^t)]^{1/t}.$$

- How find the diagonal entries of B_α^t?
- Consider the graph induced by the matrix B_α and a walk w over edges $(i_s, i_{s+1})_{s=0}^{t-1}$ such that $i_0 = i_t = i$ and $[B_\alpha]_{i_s,i_{s+1}} > 0$ for all s.
- The weight of this walk $\phi_\alpha(w)$ can be found as

$$\phi_\alpha(w) = \alpha^{c_w}\phi_1(w), \quad \text{where} \quad c_w \in \mathbb{Z} \quad \text{and} \quad \phi_1(w) = \prod_{s=0}^{t-1} [B_1]_{i_s,i_{s+1}}.$$

- Define a symmetric walk p' with edges $(i_{s+1}, i_s)_{s=0}^{t-1}$. Then, $[B_\alpha^t]_{i,i}$ contains the weights of both p and p' as summands. Hence,

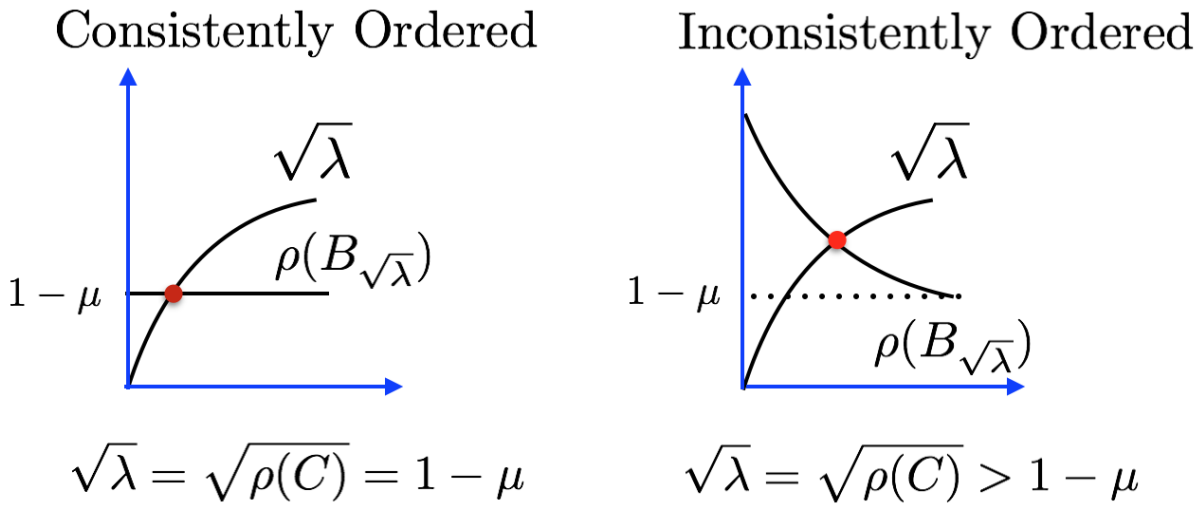
$$[B_\alpha^t]_{i,i} = \sum_{\text{all valid walks } w} \frac{\alpha^{|c_p|} + \alpha^{-|c_p|}}{2} \phi_1(w).$$

Proof Sketch of Lower Bound in Inequality (6):

- Similar to the 4×4 example, one can show $C = (I - L)^{-1} L^T \geq 0$.
- By the Perron-Frobenius Theorem, $\lambda = \rho(C)$ and $\exists z \geq 0$

$$Cz = \lambda z \quad \iff \quad (\lambda L + L^T) z = \lambda z \iff \rho(\lambda L + L^T) = \lambda$$

- Suffices to solve the equation $\rho(\lambda L + L^T) = \lambda = \sqrt{\lambda}\, \rho\left(B_{\sqrt{\lambda}}\right)$.
- We conclude by invoking Varga's lemma.



$\sqrt{\lambda} = \sqrt{\rho(C)} = 1 - \mu$ $\qquad$ $\sqrt{\lambda} = \sqrt{\rho(C)} > 1 - \mu$

Convergence Rate of CCD for M-Matrices

Corollary

Suppose Assumption 1 holds and A is an irreducible M-matrix. Then, CCD and RCD methods satisfy

$$1 < \nu_n < \frac{\mathrm{Rate(CCD)}}{\mathrm{Rate(RCD)}} \leq 2\nu_n \quad \textit{where} \quad \nu_n := \frac{\log(1-\mu)}{n\log\left(1-\frac{\mu}{n}\right)}.$$

- ν_n is a monotonically increasing function of n, where $\nu_1 = 1$ and $\lim_{n\to\infty} \nu_n = \frac{-\log(1-\mu)}{\mu} > 1$. For any $\mu \leq \frac{1}{2}$, we have $\nu_n \in [1, \frac{3}{2})$.

Corollary

Suppose Assumption 1 holds and A is an irreducible M-matrix with $n \geq 2$. Then, CCD and RCD methods satisfy $\lim_{\mu\to 0^+} \frac{\mathrm{Rate(CCD)}}{\mathrm{Rate(RCD)}} = 2$.

- CCD has a better asymptotic worst-case convergence rate than RCD.
- We quantify the amount of rate improvement and when it is achievable.

Cyclic Matrices

Definition

A matrix H is 2-cyclic if there exists a permutation matrix P such that

$$PHP^T = D + \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}, \tag{7}$$

where the diagonal null submatrices are square and D is a diagonal matrix.

- Let H be a 2-cyclic matrix that satisfy (7). Then, the graph induced by the matrix $H - D$ is periodic with period 2.
- This definition is first introduced in [Young 50], where it had an alternative name: Property A.
- It is extended to the class of p-cyclic matrices, where $p \geq 2$ in [Varga 59].
- What is the relationship between 2-cyclic matrices and consistently ordered matrices?

Lemma ([Young 71])

A matrix H is 2-cyclic if and only if there exists a permutation matrix P such that PHP^T is consistently ordered.

Convergence Rate of CCD for Cyclic Matrices

Theorem

Suppose Assumption 1 holds and A is a consistently ordered 2-cyclic matrix. Then, the spectral radius of the CCD algorithm is

$$\rho(C) = (1-\mu)^2 .$$

Corollary

Suppose Assumption 1 holds and A is a consistently ordered 2-cyclic matrix with $n \geq 2$. Then, the asymptotic worst-case rate of CCD and RCD satisfies

$$\frac{\mathrm{Rate(CCD)}}{\mathrm{Rate(RCD)}} = 2\nu_n \quad \textit{where} \quad \nu_n := \frac{\log(1-\mu)}{n\log\left(1-\frac{\mu}{n}\right)} > 1.$$

- The asymptotic worst-case convergence rate of CCD is more than 2 times faster than the one of RCD.

Numerical Experiments

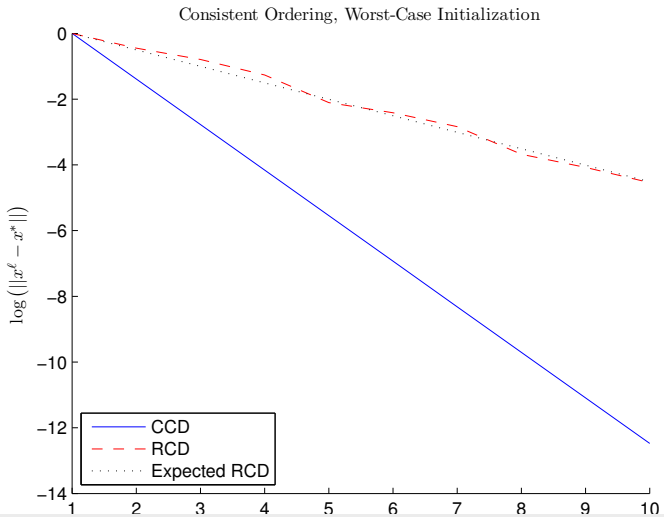
- We consider the consistently ordered 2-cyclic matrix

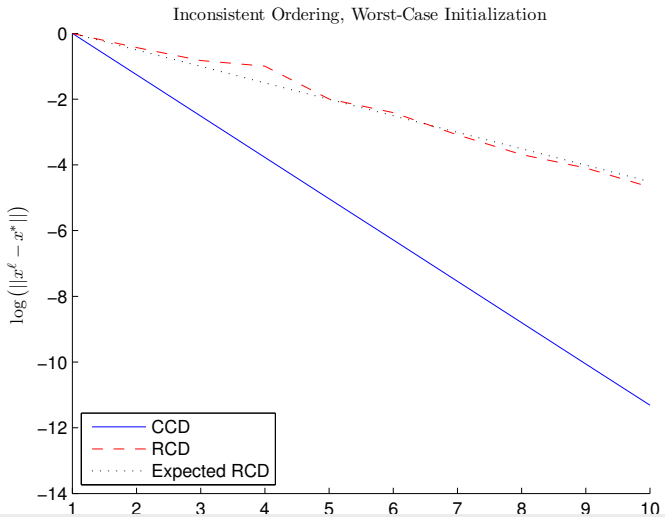
$$A = I - L - L^T, \quad \text{where} \quad L = \frac{1}{n}\begin{bmatrix} 0 & 0 \\ \mathbb{1}_{\frac{n}{2}\times\frac{n}{2}} & 0 \end{bmatrix}.$$

- For $n = 50$, the constant ν_n can be calculated as follows

$$2\nu_n = 2\frac{\log(1-\mu)}{n\log\left(1-\frac{\mu}{n}\right)} = \frac{\log(0.5)}{50\log\left(1-\frac{1}{200}\right)} \approx 2.77.$$

- Convergence to x^*. Left: Consistent ordering, Right: Inconsistent ordering.





Other related and future work

- For diagonally dominant matrices, we can show CCD is faster than RCD in a non-asymptotic sense.
- We can relax the assumption about the sign of off-diagonal entries.
- Applications:
  - Gaussian Belief Propagation: our class (M-matrices) corresponds to non-frustrated models.
  - Solving Laplacian systems, consensus.

Aggregated methods:

- Deterministic Incremental Aggregated Gradient [**M.G**., Ozdaglar, Parrilo 15]:
  - Remember past, work with delayed gradients
  - Analysis as a dynamical system with delays, we prove linear convergence.
  - Suitable for distributed optimization over networks
- Proximal Aggregated Gradient Methods [D. Vanli, Supervisors: **M.G.**, Ozdaglar.
  - Rate dependy linearly on the condition number and m.

Conclusions

- We analyzed deterministic incremental algorithms for solving additive convex cost optimization problems under smoothness assumptions.
  - We presented new rate results for a variety of stepsize rules and arbitrary orders.
- We used these results to study the random reshuffling method and presented the first analytical results for its convergence rate, which is faster than SGD.
- We provided problem classes for which CCD (or CD with any deterministic order) is faster than RCD.
- We provide a family of examples for which CCD is asymptotically faster than RCD by a factor of at least two for any dimension n.
- We provided a characterization of the best deterministic order (that leads to the maximum improvement in convergence rate).
- For diagonally dominant A, we can get similar **non-asymptotic** results.
- **Reference:** *When Cyclic Coordinate Descent Beats Randomized Coordinate Descent* (joint work with D. Vanli and A. Ozdaglar), Submitted.

Appendix

Proof of Upper Bound

- Using the same Perron-Frobenius argument,

$$(\lambda L + L^T)z = \lambda z \quad \Rightarrow \quad \lambda z^T L z + z^T L^T z = \lambda,$$

since $||z|| = 1$. Defining $\beta = z^T L z = z^T L^T z$, we get

$$\lambda = \frac{\beta}{1-\beta}. \tag{8}$$

- Since $\rho(L + L^T) = \rho(I - A) = 1 - \mu$, then for any $||y|| = 1$, we have

$$y^T(L + L^T)y \leq 1 - \mu.$$

- Picking $y = z$ yields $2\beta \leq 1 - \mu$. Using this in (8), we get

$$\lambda \leq \frac{1-\mu}{1+\mu}.$$

Conclusions

- We analyzed deterministic incremental algorithms for solving additive convex cost optimization problems under smoothness assumptions.
  - We presented new rate results for a variety of stepsize rules and arbitrary orders.
- We used these results to study the random reshuffling method and presented the first analytical results for its convergence rate, which is faster than SGD.
- We also analyzed deterministic incremental aggregated gradient and presented a new explicit linear rate result.
- Fertile research area with a significant impact in various application domains including large-scale networks and data processing.

Thank You!

Convergence Mechanism – I

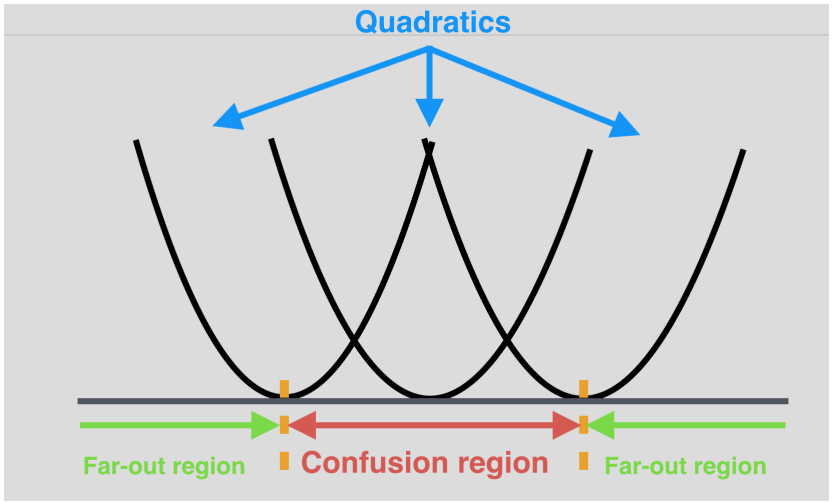


Figure: Illustration with one-dimensional quadratics [Bertsekas 15].

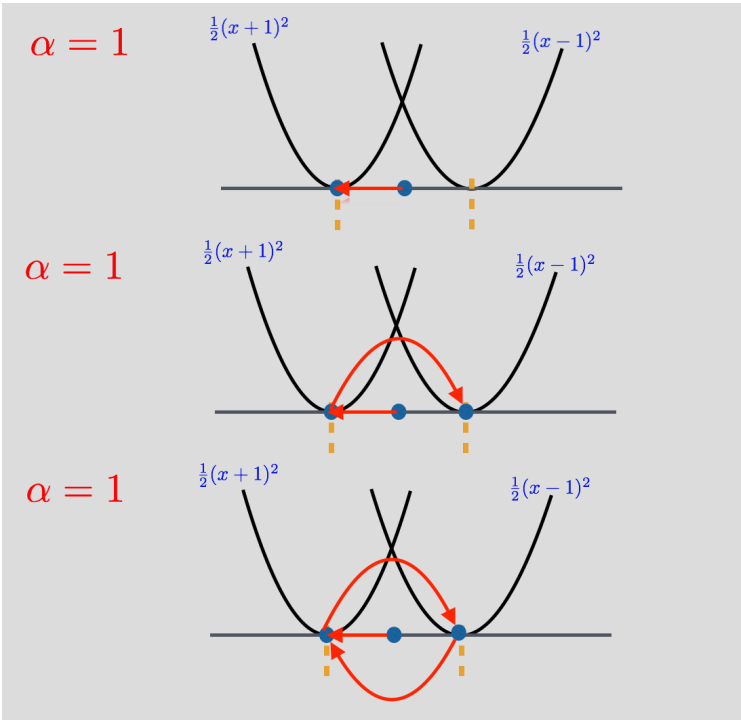
- Farout region: All individual gradients are almost as effective as the full gradient, pointing out in the right direction.
- Confusion region: Gradients are not aligned, oscillations arise.

Convergence Mechanism – II

- The choice of stepsize α_k plays an important role in the performance of incremental methods.
- A decaying stepsize is essential for global convergence to an optimal solution of the global objective function $f(x)$ [Luo 91]:

$$\alpha_k \to 0, \qquad \sum_k \alpha_k = \infty.$$

- A constant (small) stepsize ensures convergence to a neighborhood of the optimal solution [Solodov 98], [Nedic, Bertsekas 00].
  - Iterates may converge to a limit cycle [Kohonen 74].



Analysis of Incremental Gradient – I

- We analyze the method as a gradient method with error:

$$x_1^{k+1} = x_1^k - \alpha_k \big(\nabla f(x_1^k) - e^k\big),$$

$$e^k = \sum_{i=1}^{m} \big(\nabla f_i(x_1^k) - \nabla f_i(x_i^k)\big).$$

- Using smoothness, we replace $\nabla f(x_1^k) = A_k(x_1^k - x^*)$, where $A_k = \int_0^1 \nabla^2 f(x^* + \tau(x_1^k - x^*))d\tau$, and write for $\text{dist}_k = \|x_1^k - x^*\|$,

$$\text{dist}_{k+1} \leq \|\mathrm{I} - \alpha_k A_k\| \text{dist}_k + \alpha_k \|e_k\|.$$

- We use gradient Lipschitzness and boundedness to control gradient error

$$\|e^k\| \leq \alpha_k LmG.$$

- Using strong convexity bound and $\alpha_k = \frac{R}{k}$, we have for $k \geq RL$,

$$\text{dist}_{k+1} \leq \left\|\mathrm{I} - \frac{Rc}{k}\right\| \text{dist}_k + \frac{LmGR^2}{k^2}.$$

Analysis of Incremental Gradient – II

Lemma (Chung 53, Polyak 87)

Let $u_k \geq 0$ be a sequence of real numbers. Assume there exists k_0 such that

$$u_{k+1} \leq \left(1 - \frac{a}{k}\right) u_k + \frac{d}{k^{s+1}}, \quad \forall k \geq k_0,$$

where $d > 0$, $a > 0$ and $s > 0$ are real scalars. Then,

$$\begin{aligned} u_k &\leq d(a-s)^{-1} k^{-s} + o(k^{-s}) && \text{for} \quad a > s \\ u_k &= \mathcal{O}(k^{-a}) && \text{for} \quad a < s. \end{aligned}$$

- For $s = 1$, the recursion can be approximated as

$$u_{k+1} = \prod_{l=1}^{k} \left(1 - \frac{a}{l}\right) u_1 + \sum_{j=1}^{k} \left[\prod_{l=j+1}^{k-1} \left(1 - \frac{a}{l}\right) \right] \frac{d}{j^2}.$$

$$u_k \approx \underbrace{\frac{1}{k^a} u_1}_{\text{transient term}} + \underbrace{\frac{d}{a-1} \frac{1}{k}}_{\text{accumulated error}}.$$

Incremental (Sub)Gradient method

$$\begin{aligned} \min_x \quad & f(x) = \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n. \end{aligned}$$

- Idea: Sequentially take steps along the (sub)gradients of the component functions f_i.
- Each (outer) iteration consists of a cycle with m subiterations: For $k \geq 1$,

$$x_{i+1}^k = x_i^k - \alpha_k g_i^k, \quad \text{for} \quad i = 1, 2, \ldots, m,$$

  where $g_i^k \in \partial f_i(x_i^k)$ is a subgradient of f_i at x_i^k, and α_k is a stepsize.
- Outer iteration: $x_1^{k+1} := x_{m+1}^k = x_1^k - \alpha_k \sum_{i=1}^m g_i^k$.

$$\underbrace{x_1^k \quad x_2^k \quad x_3^k \quad \cdots \quad x_m^k}_{\text{one cycle}} \quad x_{m+1}^k \cdots$$

x_1^k $\qquad\qquad$ x_1^{k+1}