Randomness in Coordinate Descent

Stephen Wright

University of Wisconsin-Madison

WISO, Durham, February 2017

Outline

- Coordinate descent (CD). Cyclic and random variants.
- Known convergence results for cyclic and random cases, especially for convex quadratics.
- Computations, focusing on those that show difference between cyclic and randomized variants.
- Convergence Analysis of the "random permutations" version of CD for a convex quadratic with a permutation-invariant Hessian.
- Generalizing the permutation-invariant Hessian.

+Ching-pei Lee

Coordinate Descent (CD) Framework

... for smooth unconstrained minimization: $\min_x f(x)$:

Choose $x^0 \in \mathbb{R}^n$;
for $\ell = 0, 1, 2, \ldots$ **do**
  for $j = 0, 1, 2, \ldots, n-1$ **do**
    Define $k = \ell n + j$;
    Choose index $i = i(\ell, j) \in \{1, 2, \ldots, n\}$;
    Choose $\alpha_k > 0$;
    $x^{k+1} \leftarrow x^k - \alpha_k \nabla_i f(x^k) e_i$;
  end for
end for

- $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$: the ith coordinate vector;
- $\nabla_i f(x) = i$th component of the gradient $\nabla f(x)$;
- $\alpha_k > 0$ is the step length.

Here ℓ is the epoch, k counts individual iterations.

CD Variants: Cyclic and Random

One important differentiator between variants of CD is in the selection of index $i(\ell, j)$.

- CCD (Cyclic CD): $i(\ell, j) = j + 1$. Cycle through indices in order.
- RCD (Randomized CD): choose $i(\ell, j)$ uniformly at random from $\{1, 2, \ldots, n\}$. Sampling with replacement.
- RPCD (Random-Permutations Cyclic CD): At the start of epoch ℓ, choose a random permutation of $\{1, 2, \ldots, n\}$, denoted by $\pi_{\ell+1}$. Then $i(\ell, j)$ is the $(j+1)$th entry in $\pi_{\ell+1}$. Sampling without replacement within each cycle.

Other differentiators between CD variants:

- Choice of step $\alpha_k > 0$;
- Strategy for choosing blocks to update rather than individual indices $i(\ell, j)$;
- Parallel strategies. (M. Jordan's talk yesterday.)

Extension: Separable Regularization

$$\min_x F(x) := f(x) + \lambda\Omega(x) \tag{1}$$

- $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable;
- $\Omega(\cdot) : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function;
- $\Omega(x)$ is separable: $\Omega(x) = \sum_{i=1}^n \Omega_i(x_i)$, $\Omega_i(\cdot) : \mathbb{R} \mapsto \mathbb{R} \cup \{+\infty\}$.

Instances of $\Omega(x)$:

- Box-constrained: $\Omega(x) = \sum_{i=1}^n \mathbf{1}_{[a_i,b_i]}(x_i)$ (indicator functions);
- ℓ_1: $\Omega(x) = \|x\|_1$ or ℓ_2: $\Omega(x) = \|x\|_2^2$.
- Componentwise-nonconvex regularization is useful too: MCP or SCAD.

At iteration k, with component $i = i(\ell, j)$, obtain new iterate from

$$x_i^{k+1} := \arg\min_\zeta \nabla_i f(x^k)^T[\zeta - x_i^k] + \frac{1}{2\alpha_k}[\zeta - x_i^k]^2 + \lambda\Omega_i(\zeta).$$

If the regularization term is dropped, it reduces to the usual CD step

$$x_i^{k+1} = x_i^k - \alpha_k \nabla_i f(x^k).$$

Separable Regularization

Common to express this update using a prox-operator:

$$x_{i_k}^{k+1} = \text{prox}_{\alpha_k \lambda \Omega_{i_k}} \left(x_{i_k}^k - \alpha_k \nabla_{i_k} f(x^k) \right),$$

where

$$\text{prox}_\phi(t) := \arg\min_\zeta \frac{1}{2}\|\zeta - t\|^2 + \phi(\zeta).$$

Many interesting recent applications are regularized.

- Least squares with $\|\cdot\|_1$ regularization (LASSO) and $\|\cdot\|_2^2$ (ridge regression).
- Logistic regression regularized with $\|\cdot\|_1$.
- Support vector machines: $\|\cdot\|_2^2$ or $\|\cdot\|_1$.

Most convergence results for smooth objectives can be extended to the separable regularized case. Proof techniques are a little different, complexity results sometimes a little weaker.

Applications: Coordinate Descent

From the literature 1990-2014:

- Support vector machines (dual formulation, with kernel).
- Positron emission tomography, optical diffusion tomography.
- Protein structure - adjusting dihedral angles in a protein chain so that the end of the chain is in a specified position.
- Gene expression studies (via logistic regression).
- Recovering origin-destination matrices from traffic observations.
- Functional MRI image analysis.
- Generalized linear models in statistics.
- Transceiver design via tensor optimization.
- Phase retrieval in X-ray crystallography.
- Self-calibrating sensing models: $y = A(\theta)x$.

Why CD? Aren't Full Gradients Cheap?

Can use computational differentiation to compute a gradient of $f(x)$ at only a small constant multiple (e.g. 5) of the cost of calculating the function $f(x)$. e.g. [Griewank and Walther, 2008].

Q: Why not use full-gradient methods instead of CD?

A: Calculating $f(x)$ is too expensive for some problems, particularly when large data sets are involved. Thus $\nabla f(x)$ is also too expensive to compute fresh for each x.

However, information about ∇f can be computed or maintained much more cheaply in certain circumstances, e.g. when x changes in only one component.

Thus, each step of CD can be cheaper than a full-gradient step, so the overall cost of the algorithm may be lower.

CD is also easy to implement in a parallel asynchronous setting.

Data Analysis Example: ERM / Regression / Classification

[Nesterov, 2012]

$$f(x) = \frac{1}{m}\sum_{j=1}^{m} h_j(A_{j\cdot}x) + \lambda \sum_{i=1}^{n} \Omega_i(x_i),$$

where $A_{j\cdot}$ is j-th row of an $m \times n$ matrix A ("the data").

Maintain $g = Ax$ and $\nabla h_j(g_j)$ for $j = 1, 2, \ldots, m$. A CD step on coordinate i_k proceeds as follows:

- Compute the i_k element of the gradient of the summation term:

$$\frac{1}{m} \sum_{j=1,2,\ldots,m: A_{j,i_k} \neq 0} A_{j,i_k} \nabla h_j(g_j);$$

- Use this information, along with $\lambda \Omega_{i_k}$, to update: $x_{i_k} \leftarrow x_{i_k} + d_{i_k}$;
- Update $g_j \leftarrow g_j + A_{j,i_k} d_{i_k}$ and $\nabla h_j(g_j)$ (only for j s.t. $A_{j,i_k} \neq 0$);

Cost: $O(\text{nonzeros in } A_{\cdot i_k})$, vs $O(\text{nonzeros in } A)$ for a full-gradient method.

Convergence

- For min $f(x)$ with f smooth.
- Focus on results for f *strongly convex*:

$$f(z) \geq f(y) + \nabla f(y)^T (z - y) + \frac{\mu}{2} \|z - y\|^2.$$

- (There are also results for general convex f, and problems with regularization terms.)

Actually require a weaker condition than strong convexity:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*).$$

for some $\mu > 0$, where f^* is the optimal value. Goes by various names: "strong property," "Polyak-Łojasiewicz," "essential strong convexity." This bound holds for

- strongly convex f, with $\mu =$ modulus of convexity;
- quadratic convex f, with $\mu =$ smallest nonzero eigenvalue.

Convergence: Properties of f

Besides modulus of convexity μ and dimension n, convergence results depend on several other properties of f.

Lipschitz constant L:

$$\|\nabla f(x+v) - \nabla f(x)\| \leq L\|v\|, \quad \text{for all } x, v.$$

Componentwise Lipschitz constants:

$$|\nabla_i f(x+te_i) - \nabla_i f(x)| \leq L_i|t|, \quad i = 1, 2, \ldots, n.$$

Also

$$L_{\max} = \max_{i=1,2,\ldots,n} L_i.$$

Randomized CD (RCD) Convergence

[Nesterov, 2012] For an individual iteration, get

$$\mathbb{E}[f(x^{k+1}) - f^*] \leq \left(1 - \frac{\mu}{nL_{\max}}\right) \mathbb{E}[f(x^k) - f^*], \quad k = 0, 1, 2, \ldots,$$

where expectation is taken over the random indices $i(\ell, j)$ chosen i.i.d. at each iteration.

For an epoch — from $x^{\ell n}$ to $x^{(\ell+1)n}$ — get rate:

$$\left(1 - \frac{\mu}{nL_{\max}}\right)^n \approx \left(1 - \frac{\mu}{L_{\max}}\right).$$

Thus to get a factor-of-ϵ decrease in expected f error, need about

$$\frac{L_{\max}}{\mu} |\log \epsilon| \quad \text{epochs of RCD.}$$

[Nesterov, 2012] also has an R-linear result with rate twice as fast: approx $1 - 2\mu/L_{\max}$ factor per epoch.

Rate Comparison: RCD vs. (Full)-Gradient Descent

Full-gradient steepest descent: $x^{\ell+1} = x^\ell - \alpha_\ell \nabla f(x^\ell)$,

For strongly convex f with steplength $\alpha_\ell \equiv 1/L$, get linear convergence:

$$f(x^{\ell+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)[f(x^\ell) - f^*],$$

where L is the Lipschitz constant for the full gradient: Thus get factor-of-ϵ reduction in f error in

$$\frac{L}{\mu}|\log \epsilon| \quad \text{steepest-descent iterations.}$$

If we assume that n steps of RCD (one **epoch**) costs about the same as one step of steepest descent, can compare directly with the $(L_{\max}/\mu)|\log \epsilon|$ rate promised by RCD: factor of about $L/L_{\max}$ improvement for RCD.

Lipschitz constant ratio:

$$1 \leq L/L_{\max} \leq n.$$

(Upper bound achieved by $f(x) = x^T \mathbf{1}\mathbf{1}^T x$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$.)

Cyclic CD (CCD) Convergence

Cycle through components in fixed order:

$$1, 2, 3, \ldots, n, 1, 2, 3, \ldots, n, 1, 2, 3, \ldots .$$

Analysis of [Beck and Tetruashvili, 2013] treats CCD as an approximate form of Steepest Descent, bounding improvement in f over one cycle in terms of the gradient at the start of the cycle:

$$f(x^{(\ell+1)n}) \leq f(x^{\ell n}) - \eta \|\nabla f(x^{\ell n})\|^2, \quad \text{for some } \eta > 0.$$

For $\alpha_k \equiv \alpha \leq 1/L_{\max}$, get linear convergence:

$$[f(x^{(\ell+1)n}) - f^*] \leq \left[1 - \frac{\mu}{(2/\alpha)(1 + nL^2\alpha^2)}\right] [f(x^{\ell n}) - f^*], \quad \ell = 0, 1, 2, \ldots .$$

Paradoxically, rate for exact line search at each step is even slower.

Compare Rates: CCD vs RCD vs Steepest Descent

Different choices of α in CCD yield different complexity estimates for factor-of-ϵ improvement:

$$\text{CCD with } \alpha = 1/L_{\max}: \quad \frac{2nL^2/L_{\max}}{\mu}|\log\epsilon|$$

$$\text{CCD with } \alpha = 1/L: \quad \frac{2Ln}{\mu}|\log\epsilon|$$

$$\text{CCD with } \alpha = 1/(\sqrt{n}L): \quad \frac{4L\sqrt{n}}{\mu}|\log\epsilon|.$$

- Worse than expected linear rate for RCD by factors of

$$\sqrt{n}\frac{L}{L_{\max}} \text{ to } n\frac{L^2}{L_{\max}^2};$$

- Worse than linear rate for steepest descent by factors of

$$\sqrt{n} \text{ to } n\frac{L}{L_{\max}}.$$

CCD with Exact Search (on Convex Quadratics)

[Sun and Ye, 2016] have improved results for CCD with exact line search on *convex quadratic* f.

Per-epoch rate is bounded above by

$$\rho = 1 - \max\left\{ \frac{\mu L_{\min}}{n L L_{\text{avg}}}, \frac{\mu L_{\min}}{L^2(2 + \log n/\pi)^2}, \frac{\mu L_{\min}}{n^2 L_{\text{avg}}^2} \right\},$$

where $L_{\min}$ and L_{avg} are minimum and average values of the diagonals of the Hessian.

Return to this later, when we consider a specific matrix.

Random-Permutations CD (RPCD)

Can apply the CCD theory directly, so same worst-case rates apply to RPCD and CCD.

But computational behavior tracks RCD much more closely than CCD!

Is there a theory that explains differences between CCD and RPCD?

Also understand the different rates for CCD, RCD, Steepest Descent.

- Does practical behavior track worst-case bounds?
- If so, for what kinds of problems?

Convex Quadratics

Even for the simplest class of convex nonlinear problems — convex quadratics — the behavior of CD and gradient methods is nontrivial to analyze.

Perhaps not surprising: Methods that use acceleration / momentum, such as Conjugate Gradient and Nesterov's accelerated gradient, are also nontrivial to analyze on convex quadratics.

Understanding behavior on convex quadratics is a good step toward understanding general behavior.

But it's hardly the whole story, otherwise nonlinear conjugate gradient would much better than it does in practice!

CD and Gauss-Seidel

$$f(x) = \frac{1}{2} x^T A x, \quad \text{where } A \text{ is symmetric positive semidefinite.}$$

Solution $x^* = 0$ with optimal objective $f^* = 0$. Nonzero initial point x^0.

It satisfies the "strong" / PL property, even when singular, so linear convergence of $f(x^k) - f^*$ is always observed, where μ is the smallest *nonzero* eigenvalue of A.

CCD with exact line search on this function is exactly Gauss-Seidel for $Ax = 0$.

RPCD and RCD are forms of randomized Gauss-Seidel.

Successive over-relaxation (SOR) correspond to CD in which we overshoot the exact minimizing α_k by some fixed fraction at each iteration k.

Gauss-Seidel / CCD

Write $A = L + D + L^T$, where L is strictly lower triangular, D is diagonal.

One cycle of Gauss-Seidel / Cyclic CD satisfies:

$$x^+ = -(L+D)^{-1}(L^T x) = Cx, \quad \text{where } C := -(L+D)^{-1}L^T.$$

After ℓ cycles, we have

$$f(x^{\ell n}) = \frac{1}{2}(x^0)^T (C^T)^\ell A C^\ell x^0, \quad \mathbb{E}_{x^0} f(x^{\ell n}) = \text{trace}\left((C^T)^\ell A C^\ell\right).$$

assuming $x^0 \sim N(0, I)$.

Jordan decomposition $C = RJR^{-1}$, where J contains Jordan blocks of constructed from eigenvalues γ_i, $i = 1, 2, \ldots, n$, with $|\gamma_i| < 1$ since A is positive definite. Obtain

$$(R^{-1}x^{\ell n}) = J^\ell (R^{-1}x^0), \quad \text{with } J^\ell \to 0 \text{ as } \ell \to \infty.$$

Gelfand's formula: average reduction in $\|x\|$ per cycle is $\rho(C)$.
Suggests asymptotic per-epoch rate of $\rho(C)^2$ for $f(x^{\ell n})$.

Randomized Gauss-Seidel / RPCD

Randomized GS, with permutation matrix P_ℓ on cycle ℓ:

$$P_\ell^T A P_\ell = L_\ell + D_\ell + L_\ell^T, \quad \text{where } C_\ell := -(L_\ell + D_\ell)^{-1} L_\ell^T.$$

One cycle is $x^+ \leftarrow P_\ell C_\ell P_\ell^T \hat{x}$, so we have

$$x^{\ell n} = P_\ell C_\ell P_\ell^T P_{\ell-1} C_{\ell-1} P_{\ell-1}^T \ldots P_1 C_1 P_1^T x^0.$$

Question: Can we evaluate

$$\mathbb{E}_{P_\ell, P_{\ell-1}, \ldots, P_1, x^0} \|x^{\ell n}\|^2 \quad \text{or} \quad \mathbb{E}_{P_\ell, P_{\ell-1}, \ldots, P_1, x^0} f(x^{\ell n})$$

and find matrices for which it is provably better than the values obtained through CCD?

YES! (for some matrices). See later....

Computations with Quadratics

Define

$$A = V\Sigma V^T + \zeta \mathbf{1}\mathbf{1}^T,$$

where

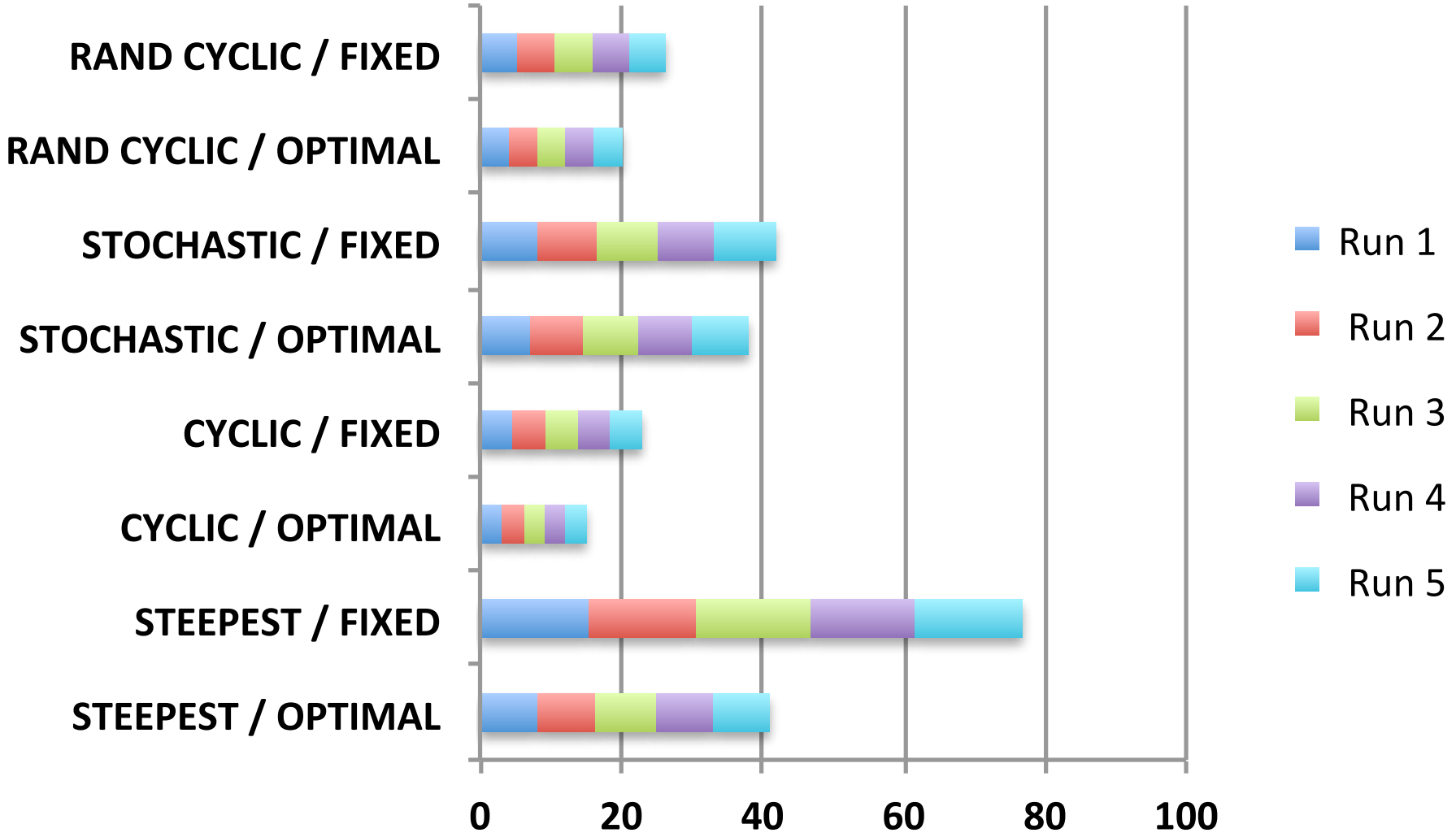
- V is a random orthogonal;
- Σ is a positive diagonal, where Σ_{ii} follow a log-uniform distribution. Specifically: $\log_{10} \Sigma_{ii} \sim U[-\sigma, 0]$ for $\sigma = 1, 2, 3, 4$.
- $\zeta > 0$.

For convenience scale A so that $L_{\max} = 1$, and start from $x^0 \sim N(0, I)$.

What does convergence theory suggest?

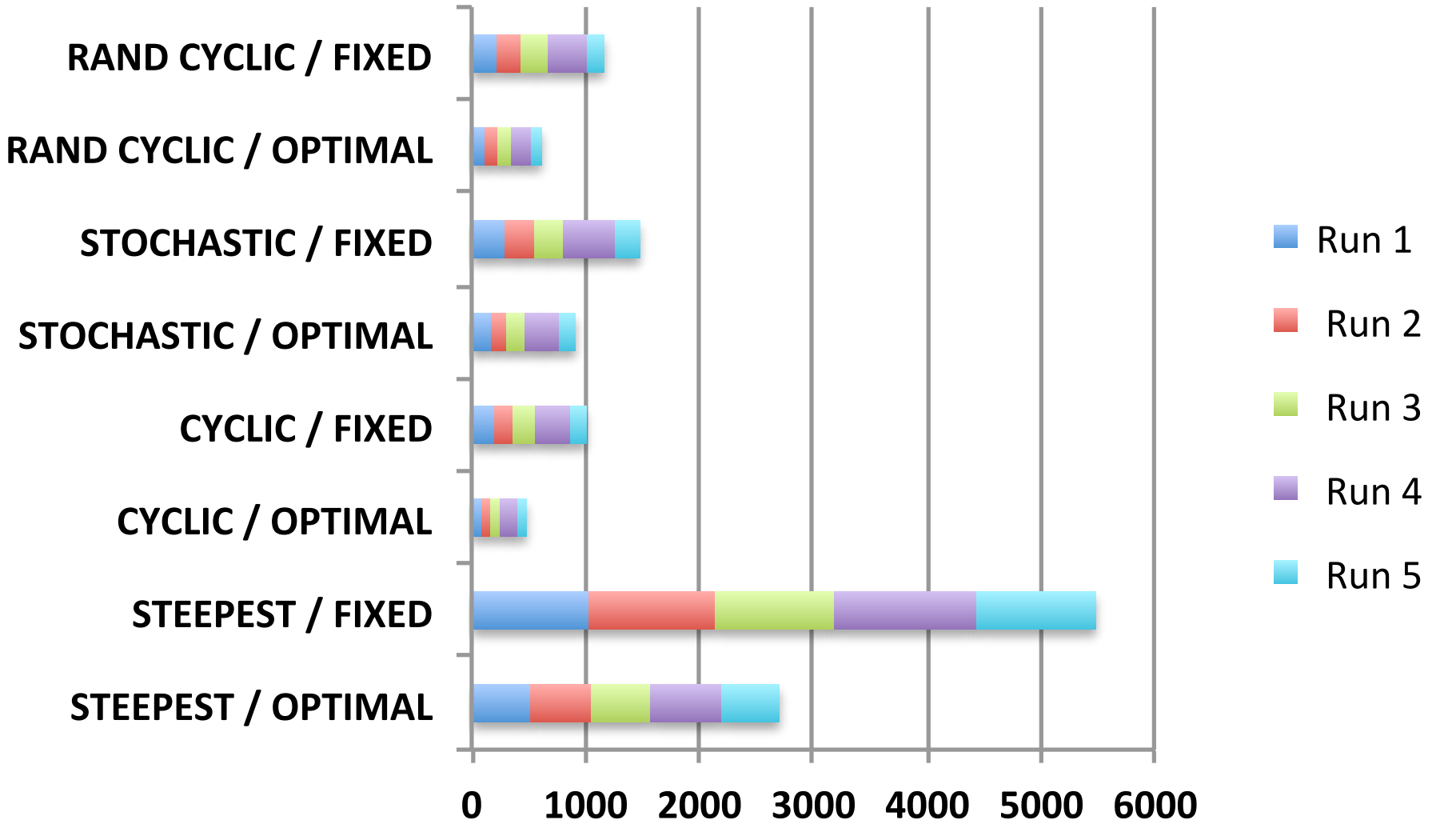
- $\zeta = 0$ has $L/L_{\max}$ modest, so no particular advantage for CD methods over Steepest Descent. All methods should degrade as condition number increases (i.e. σ increases).
- Larger ζ means that one eigenvalue dominates: $L/L_{\max}$ is large. Theory suggests a good case for random CD methods.

Computations: Log-Uniform Eigenvalues: $\text{cond}(A) \approx 10$



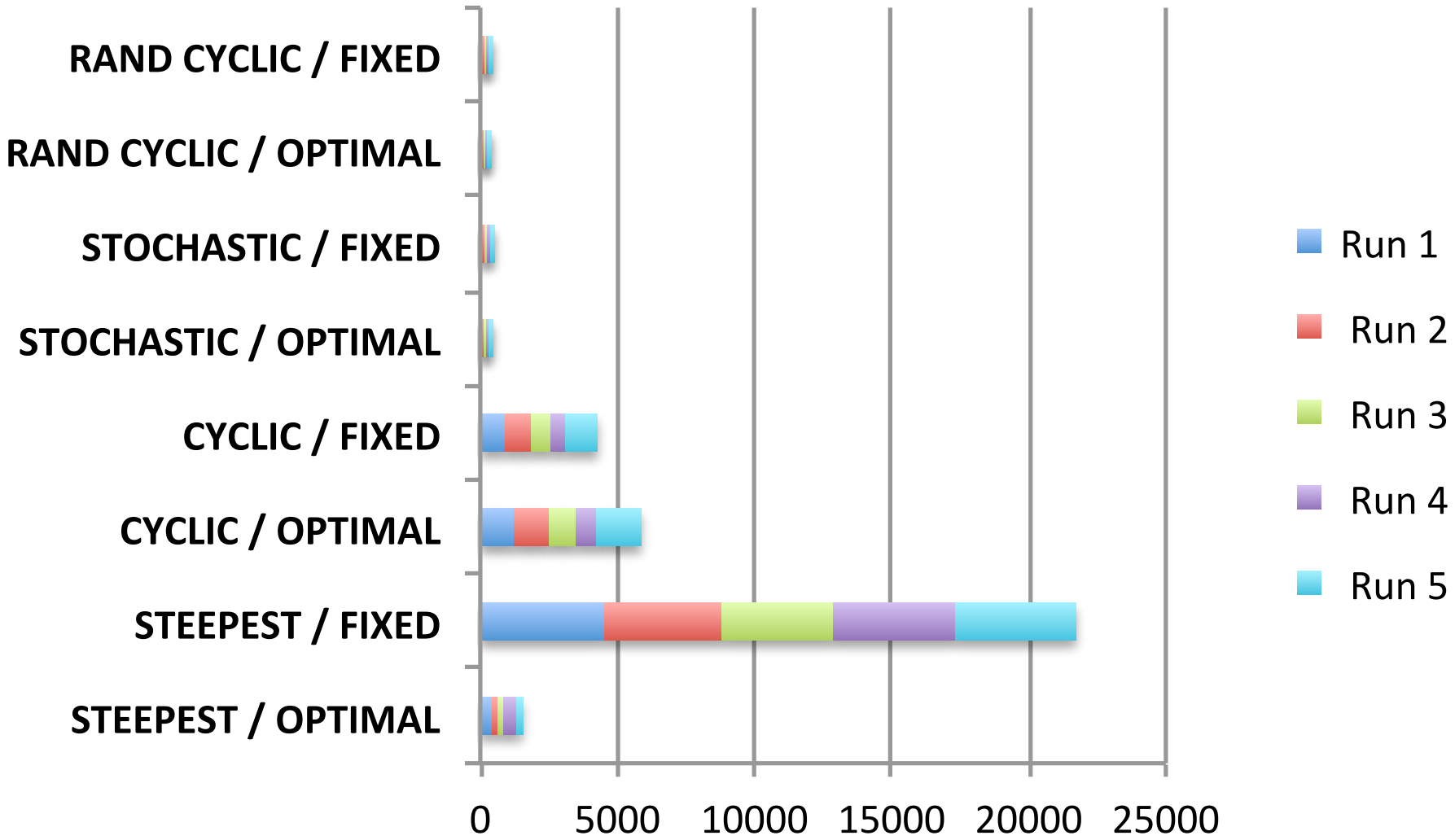
$L/L_{\max} \approx 2.$

Computations: Log-Uniform Eigenvalues: cond(A) ≈ 850



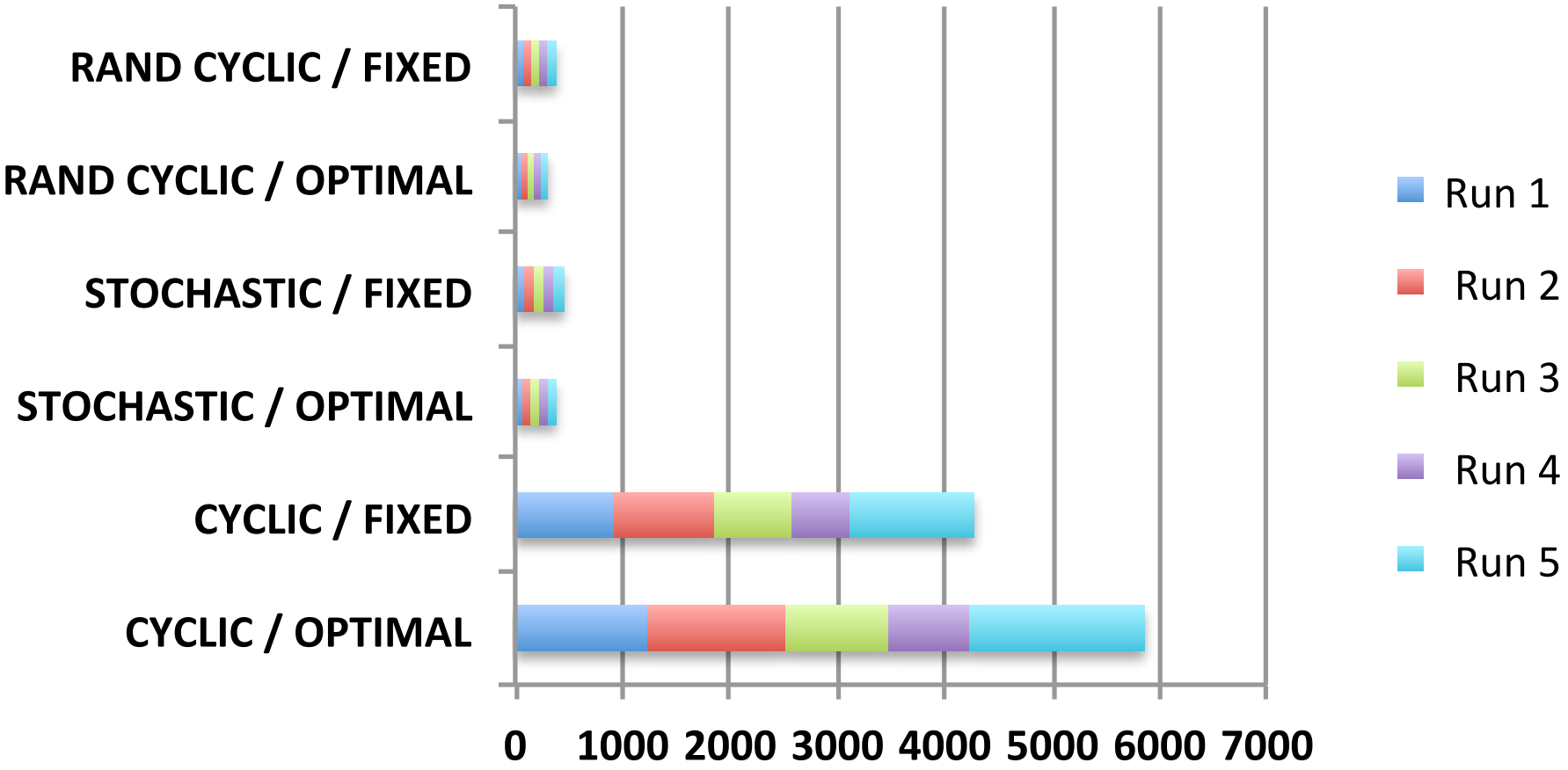
$L/L_{\max} \approx 4$.

Computations: Large $\mathbf{1}\mathbf{1}^T$ component in A ($\zeta = .3$)



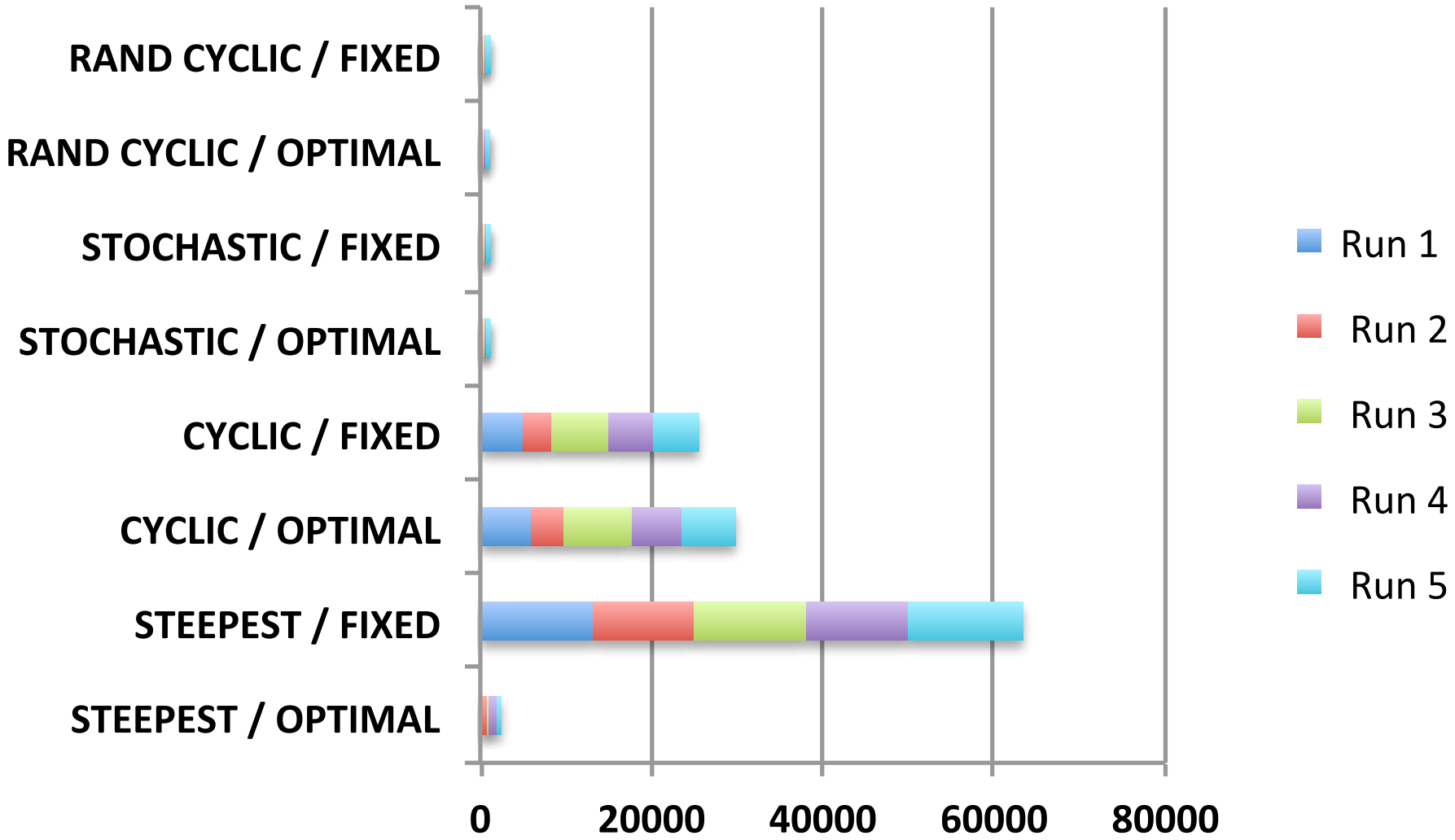
$\text{cond}(A) \approx 3000$, $L/L_{\max} \approx 50$, $\zeta = 0.3$.
Steepest Descent / Fixed and Cyclic CD are poor.

Computations: (Same, with Steepest omitted)



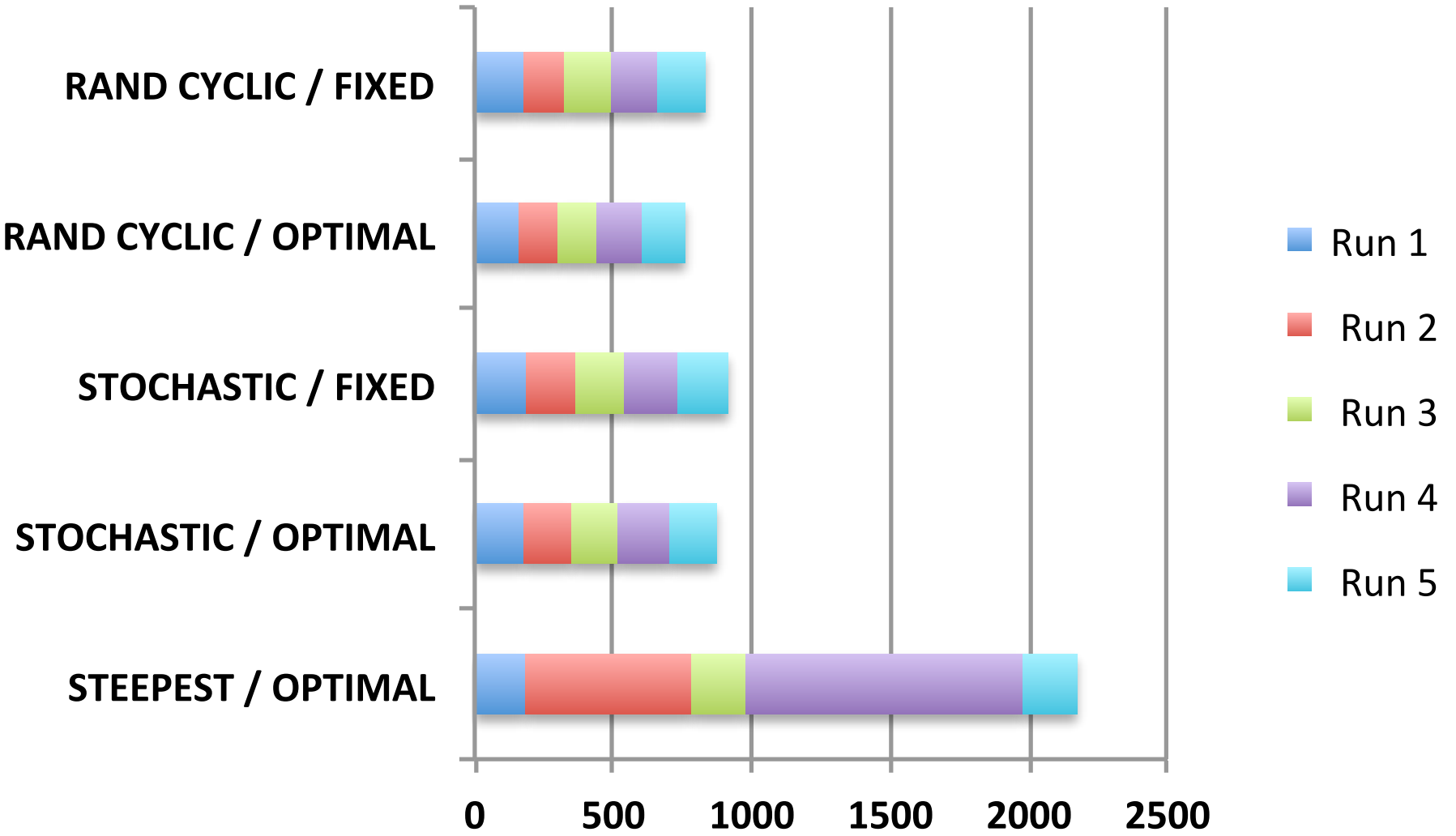
Stochastic and Random Cyclic are good, with not much difference between optimal and fixed steps.

Computations: Even Larger $\mathbf{1}\mathbf{1}^T$ component in A ($\zeta = 1$)



$\text{cond}(A) \approx 10,000$, $L/L_{\max} \approx 75$, $\zeta = 1$.
Again, Steepest Descent / Fixed and Cyclic CD are poor.

Computations: (Same, but only the good ones)



Stochastic and Randomized Cyclic remarkably similar! Optimal step makes little difference. Steepest / Optimal is competitive—Why?

Permutation-Invariant A

Define A to be a symmetric matrix that is

- Invariant under symmetric permutations: $P^T AP = A$ for all P;
- 1s on the diagonal;
- Positive semidefinite.

These conditions define the following class of matrices:

$$A = \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T, \text{where } \delta \in [0,1), \tag{2}$$

where $\mathbf{1} = (1,1,\ldots,1)^T$, whose eigenvalues are

$$\delta + n(1-\delta), \delta, \delta, \ldots, \delta.$$

$\delta > 0$ small $\Rightarrow$ one dominant eigenvalue. RPCD is much faster than CCD.

Permutation invariance means that $C_\ell \equiv C = -(L+D)^{-1}L^T$ for all ℓ, so

$$x^{\ell n} = P_\ell C P_\ell^T P_{\ell-1} C P_{\ell-1}^T \ldots P_1 C P_1^T x^0.$$

(For CCD, have $x^{\ell n} = C^\ell x^0$.)

RCD and CCD: Per-Epoch Convergence

Use results from [Sun and Ye, 2016] to get upper and lower bounds on per-epoch convergence for **CCD**:

$$\rho_{\text{CCD}}(\delta, x^0) \leq 1 - \frac{\delta}{n(n(1-\delta)+\delta)}.$$

$$\rho_{\text{CCD}}(\delta, x^0) \geq \left(1 - \frac{2\delta\pi^2}{n(n(1-\delta)+\delta)}\right)^2.$$

Both have the form $1 - c\delta/n^2$, where c is a modest quantity.

For RCD, the expected per-epoch improvement is

$$\rho_{\text{RCD}}(\delta, \text{predicted}) \approx 1 - \delta. \tag{3}$$

Thus iteration complexity of RCD is $O(n^2)$ better than CCD.

RPCD Convergence

Function value after ℓ epochs is

$$f(x^{\ell n}_{\text{RPCD}}) = \frac{1}{2}(x^0)^T \left(P_1 C^T P_1^T \ldots P_\ell C^T P_\ell^T A P_\ell C P_\ell^T \ldots P_1 C P_1^T \right) x^0.$$

Take expectation w.r.t. $P_1, P_2, \ldots, P_\ell$, and then w.r.t. x^0.

Expectation w.r.t. x^0 is easy when we assume that components of x^0 are i.i.d from $N(0,1)$ — it's just the trace of the matrix.

Note that $P_1, P_2, \ldots, P_\ell$ are independent permutation matrices. Can take expectations from the "inside out," starting with P_ℓ, then $P_{\ell-1}, \ldots, P_1$. Define $\bar{A}^{(t)}$, $t = 0, 1, 2, \ldots, \ell$ as:

$$\bar{A}^{(t)} = \mathbb{E}_{P_{\ell-t+1},\ldots,P_\ell} \left(P_{\ell-t+1} C^T P_{\ell-t+1}^T \ldots P_\ell C^T P_\ell^T A P_\ell C P_\ell^T \ldots P_{\ell-t+1} C P_{\ell-t+1}^T \right),$$

and note that $\bar{A}^{(0)} = A$ and that

$$\mathbb{E} f(x^{\ell n}_{\text{RPCD}}) = \frac{1}{2} \mathbb{E}_{x^0} \left[(x^0)^T \bar{A}^{(\ell)} x^0 \right].$$

Key Technical Lemma

Lemma

Given any matrix $Q \in \mathbb{R}^{n \times n}$ we have

$$B := \mathbb{E}_P[PQP^T] = \tau_1 I + \tau_2 \mathbf{1}\mathbf{1}^T, \tag{4}$$

where

$$\tau_2 = \frac{\mathbf{1}^T Q \mathbf{1} - \mathrm{trace}(Q)}{n(n-1)}, \quad \tau_1 = \frac{\mathrm{trace}(Q)}{n} - \tau_2. \tag{5}$$

All off-diagonals are identical; all diagonals are identical.

Taking expectation over the permutation has a homogenizing effect.

Immediate Consequence

$$\mathbb{E}_P(P^T C^T C P) = d_1 I + d_2 \mathbf{1}\mathbf{1}^T, \tag{6a}$$

$$\mathbb{E}_P(P^T C^T \mathbf{1}\mathbf{1}^T C P) = m_1 I + m_2 \mathbf{1}\mathbf{1}^T, \tag{6b}$$

where

$$d_2 = \frac{\mathbf{1}^T C^T C \mathbf{1} - \text{trace}(C^T C)}{n(n-1)} = \frac{\|C\mathbf{1}\|_2^2 - \|C\|_F^2}{n(n-1)} \tag{7a}$$

$$d_1 = \frac{\text{trace}(C^T C)}{n} - d_2 = \frac{\|C\|_F^2}{n} - d_2 \tag{7b}$$

$$m_2 = \frac{(\mathbf{1}^T C \mathbf{1})^2 - (\mathbf{1}^T C)(C^T \mathbf{1})}{n(n-1)} = \frac{(\mathbf{1}^T C \mathbf{1})^2 - \|C^T \mathbf{1}\|_2^2}{n(n-1)} \tag{7c}$$

$$m_1 = \frac{(\mathbf{1}^T C)(C^T \mathbf{1})}{n} - m_2 = \frac{\|C^T \mathbf{1}\|_2^2}{n} - m_2 \tag{7d}$$

Estimating d_1, d_2, m_1, m_2 for the Invariant Matrix

$$
\begin{aligned}
d_1 &\approx 1 - 2\delta - 2\frac{\delta}{n} \\
m_2 &\approx \frac{2\delta^3}{n^2}, \\
d_2 &\approx 1 - 2\delta - \frac{2}{n} + 4\frac{\delta}{n}, \\
m_1 &\approx \frac{\delta^2}{n}.
\end{aligned}
$$

(Leave out higher-order terms in δ and $1/n$.)

Proof: About 4 pages of asymptotics based on (7) and the definition of $C = -(L+D)^{-1}L^T$, which has

$$
C_{ij} = \begin{cases} -(1-\delta)\delta^{i-1} & \text{for } i < j \\ (1-\delta)(\delta^{i-j} - \delta^{i-1}) & \text{for } i \geq j. \end{cases}.
$$

Two-Variable Recurrence

Given the sequence of $\bar{A}^{(t)}$ matrices defined above, we have the following relationship between two successive terms:

$$\begin{aligned}\bar{A}^{(t+1)} &= \mathbb{E}_{P_{\ell-t}}(P_{\ell-t}C^T P_{\ell-t}^T \bar{A}^{(t)} P_{\ell-t} C P_{\ell-t}^T) \\ &= \mathbb{E}_P(PC^T P^T \bar{A}^{(t)} PCP^T).\end{aligned}$$

Assume that $\bar{A}^{(t)} = \eta_t I + \nu_t \mathbf{1}\mathbf{1}^T$ — permutation invariant. (True for $t = 0$, since $\bar{A}^{(0)} = A = \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T$.) Then

$$\begin{aligned}\eta_{t+1} I + \nu_{t+1}\mathbf{1}\mathbf{1}^T &= \bar{A}^{(t+1)} \\ &= \mathbb{E}_P(PC^T P^T \bar{A}^{(t)} PCP^T) \\ &= \mathbb{E}_P(PC^T \bar{A}^{(t)} CP^T) \\ &= \eta_t \mathbb{E}_P(PC^T CP^T) + \nu_t \mathbb{E}_P(PC^T \mathbf{1}\mathbf{1}^T CP^T) \\ &= \eta_t(d_1 I + d_2 \mathbf{1}\mathbf{1}^T) + \nu_t(m_1 I + m_2 \mathbf{1}\mathbf{1}^T).\end{aligned}$$

Two-Variable Recurrence

So we have

$$\begin{bmatrix} \eta_{t+1} \\ \nu_{t+1} \end{bmatrix} = M \begin{bmatrix} \eta_t \\ \nu_t \end{bmatrix} = M^{t+1} \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix},$$

where

$$M := \begin{bmatrix} d_1 & m_1 \\ d_2 & m_2 \end{bmatrix} \approx \begin{bmatrix} 1-2\delta-2\frac{\delta}{n} & \frac{\delta^2}{n} \\ 1-2\delta-\frac{2}{n}+4\frac{\delta}{n} & \frac{2\delta^3}{n^2} \end{bmatrix}.$$

In particular,

$$\begin{bmatrix} \eta_1 \\ \nu_1 \end{bmatrix} = M \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix} \approx \begin{bmatrix} (1-2\delta)\delta \\ (1-2\delta-2/n)\delta \end{bmatrix} \approx \begin{bmatrix} \delta \\ \delta \end{bmatrix}.$$

Reduction on first epoch:

$$\mathbb{E}f(x^0) = \frac{1}{2}\mathrm{trace}(\bar{A}^{(0)}) = \frac{n}{2} \quad \rightarrow \quad \mathbb{E}f(x^1) \approx \frac{n}{2}\delta(1-2\delta),$$

a reduction factor of about $\delta(1-2\delta)$.

Later Epochs

Behavior of $\mathbb{E}f(x^{\ell n})$ on later epochs governed by eigendecomposition of M. Eigenvalues are

$$\lambda_1 = 1 - 2\delta - \frac{2\delta}{n} + O(\delta^2), \quad \lambda_2 \approx -\frac{\delta^2}{n}.$$

We find that

$$\begin{bmatrix} \eta_t \\ \nu_t \end{bmatrix} = M^t \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix} \approx \left(1 - 2\delta - \frac{2\delta}{n}\right)^t \begin{bmatrix} \delta \\ \delta \end{bmatrix}, \quad t = 1, 2, \ldots .$$

So

$$\mathbb{E}f(x^{\ell n}) \approx n(1-2\delta)^\ell \delta, \quad \ell = 1, 2, \ldots,$$

and a Q-linear rate of approximately:

$$\rho_{\text{RPCD}}(\delta) \approx (1 - 2\delta).$$

Similar to iteration complexity of RCD.

First Iteration

Large reduction is achieved on the very first iteration — most of the improvement in f in the first epoch happens here.

This happens for all variants (CCD, RCD, RPCD) provided the line search is exact. $f(x^1)$ is about $2\delta f(x^0)$.

Theorem

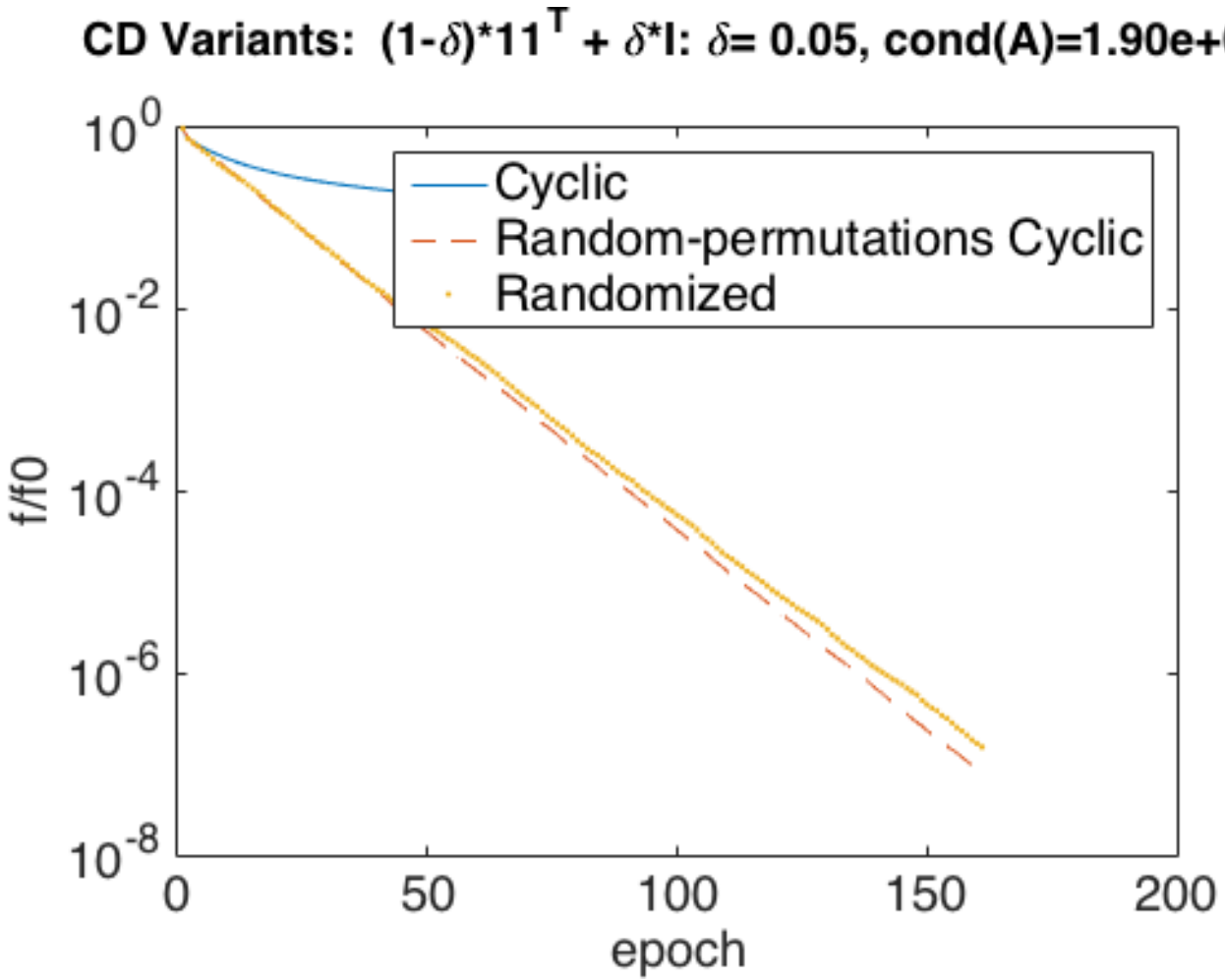
Consider RCD or RPCD with exact line search. Suppose that $x^0 \sim N(0, I)$. Then after a single iteration, the expected function value is as follows:

$$\mathbb{E}_{x^0}\mathbb{E}_i f(x^1) = \frac{(n-1)}{2}\delta(2-\delta) = \frac{n-1}{n}\delta(2-\delta)\mathbb{E}_{x^0} f(x^0), \qquad (8)$$

where i denotes the coordinate chosen for updating at the first iteration. For the CCD, the same estimate applies with no expectation w.r.t. i.

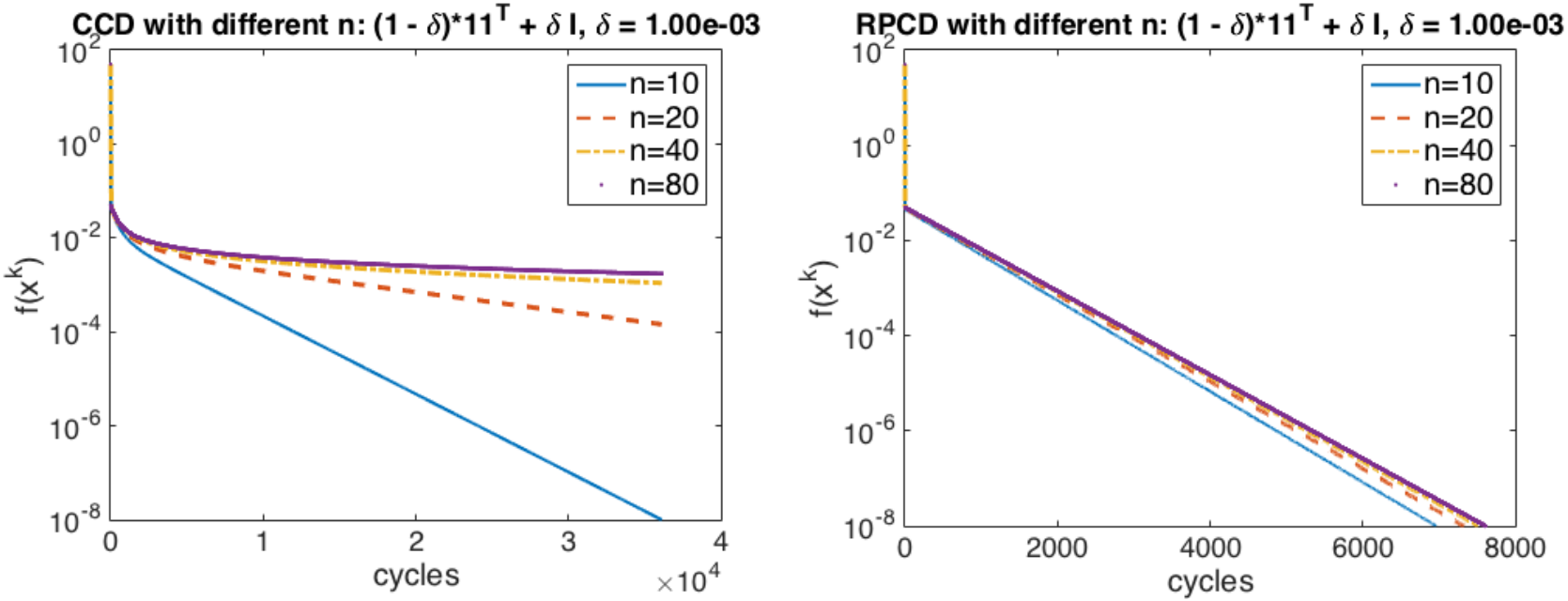
Proof: Simply use formulas and take some elementary expectations.

Computations with Invariant A



CCD, RCD, and RPCD with $n = 100$ and $\delta = .05$.

Dependence on n



Convergence of f for CCD and RPCD with $\delta = .001$ and $n = 10, 20, 40, 80$. Convergence rate of CCD deteriorates as n grows, as predicted by the estimates, while the convergence rate of RPCD is independent of n.

Extending the "Invariant A" Analysis

Computations showed that for *most* A with a single dominant eigenvalue reveal big differences between CCD and RCD/RPCD performance.

Given $\delta \in (0,1)$ (small) and $\epsilon > 0$ with $\delta = O(\epsilon)$, consider matrices

$$B_{u,\epsilon} := \delta I + (1-\delta) u u^T,$$

with

$$|u_i| \in [\sqrt{\delta/(\delta+\epsilon)}, 1], \quad i = 1, 2, \ldots, n.$$

(The u_i are confined to a limited range.) Defining $U := \mathrm{diag}(u)$, we have

$$A_\epsilon := U^{-1} B_{u,\epsilon} U^{-1} = \delta U^{-2} + (1-\delta)\mathbf{1}\mathbf{1}^T, \tag{9}$$

and the diagonals U^{-2} are in $[1, \epsilon/\delta + 1]$. Thus can write $\delta U^{-2} = \delta I + \epsilon D$, where D is diagonal with elements in $[0,1]$, so

$$A_\epsilon := \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T + \epsilon D,$$
$$D = \mathrm{diag}(d), \text{ with } d_i \in [0,1] \text{ for all } i = 1, 2, \ldots, n.$$

Off-Diagonally Invariant Matrix

Thus we focus on analyzing RPCD, CCD, and RCD on the matrix

$$A_\epsilon := \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T + \epsilon D,$$

for small positive δ and larger ϵ (but still $\epsilon \ll 1$).

Notation:

$$D = \mathrm{diag}(d_1, d_2, \ldots, d_n), \quad d_{\mathrm{av}} = \mathbf{1}^T d/n, \quad d_{\mathrm{av},2} = d^T d/n.$$

CD applied to this matrix similar to CD applied to $B_{u,\epsilon} := \delta I + (1-\delta)uu^T$ — functions differ by a factor of $\delta/(\epsilon+\delta)$ and 1.

Setup for analysis is similar to before **but** the single-epoch operator C now depends on permutation P.

We lose the property that made the "invariant A" case tractable.

But only the diagonals of A_ϵ are affected by P — this helps.

One Epoch

$$A_\epsilon := \delta I + (1-\delta)\mathbf{1}\mathbf{1}^T + \epsilon D,$$

Define

$$D_P := P^T D P, \quad E := \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \end{bmatrix}.$$

Then the single-epoch iteration matrix is

$$C_P := -(1-\delta)\big((1-\delta)E + (I + \epsilon D_P)\big)^{-1} E^T,$$

and we have

$$x^{tn} = P_t C_{P_t} P_t^T x^{(t-1)n}, \quad t = 1, 2, \dots .$$

Analysis Setup

Function value after ℓ epochs is

$$f_\epsilon(x^{\ell n}) = \frac{1}{2}(x^0)^T \left(P_1 C_{P_1}^T P_1^T \ldots P_\ell C_{P_\ell}^T P_\ell^T A_\epsilon P_\ell C_{P_\ell} P_\ell^T \ldots P_1 C_{P_1} P_1^T \right) x^0.$$

We're interested in the expectation w.r.t. $x^0 \sim N(0, I)$ and w.r.t. permutations $P_1, P_2, \ldots, P_\ell$, which are independent.

As for the invariant A, define a sequence of expectation matrices:

$$\begin{aligned}
\bar{A}_\epsilon^{(0)} &= A_\epsilon, \\
\bar{A}_\epsilon^{(1)} &= \mathbb{E}_{P_\ell} \left(P_\ell C_{P_\ell}^T P_\ell^T \right), \\
&\vdots \\
\bar{A}_\epsilon^{(\ell)} &= \mathbb{E}_{P_1,\ldots,P_\ell} \left(P_1 C_{P_1}^T P_1^T \ldots P_\ell C_{P_\ell}^T P_\ell^T A_\epsilon P_\ell C_{P_\ell} P_\ell^T \ldots P_1 C_{P_1} P_1^T \right).
\end{aligned}$$

Relationship between successive members of the sequence is

$$\bar{A}_\epsilon^{(t)} = \mathbb{E}_P (P C_P^T P^T \bar{A}_\epsilon^{(t-1)} P C_P P^T).$$

Approach

Similar to for the invariant A:

- Seek a compact representation of each $\bar{A}_\epsilon^{(t)}$, $t = 0, 1, 2, \ldots$ in terms of a few parameters;
- Derive a recurrence for these parameters from t to $t + 1$;
- Translate this recurrence into statements about decrease in expected function value over an epoch.

..... but it's a lot more complicated the invariant-A case:

- Recurrence has **seven** parameters (instead of two);
- Technical details (expectations over permutations) are a lot trickier;
- $O(\epsilon^3)$ remainder terms arise — we can prove convergence down to this level. (By adding terms to the representation, could improve this to $O(\epsilon^4)$ and beyond.)

The analysis gives the broad picture of convergence behavior, that accords with empirical evidence. Overall per-epoch rate is about $(1 - \delta)^2$, as for invariant A.

The "Compact" Representation

$$\bar{A}_\epsilon^{(t)} = \eta_t I + \nu_t \mathbf{1}\mathbf{1}^T + \epsilon_t D + (\tau_t/2)(d\mathbf{1}^T + \mathbf{1}d^T) + \beta_t D^2 + (\zeta_t/2)(D^2\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^T D^2) + \chi_t dd^T + O(\epsilon^3).$$

- Seven parameters: η_t and ν_t are the same as in the invariant A, while the remainder arise from the ϵD term.
- The form of D is preserved across epochs, but powers of D appear, and "spreading" occurs.

To derive the recurrence, need to complete the 7×7 table of showing how each of the seven terms changes in expectation over one epoch:

$$\mathbb{E}_P(PC_P^T P^T \square PC_P P^T),$$

where $\square$ is I, $\mathbf{1}\mathbf{1}^T$, D, $d\mathbf{1}^T + \mathbf{1}d^T$, ...

We save only the most significant terms, ensuring that no term of order lower than ϵ^3 is ignored.

Expected Attenuation over One Epoch

All scaled by $(1-\delta)^2$.

	I	$\mathbf{1}\mathbf{1}^T$	D	$d\mathbf{1}^T$	dd^T	D^2	$D^2\mathbf{1}\mathbf{1}^T$
I	$1-\frac{2}{n}\delta$	$1-\frac{2}{n}-\frac{2n}{n-1}\epsilon d_{\text{av}}$	$2\,\frac{n+1}{n}\epsilon$	$\frac{3n-2}{n(n-1)}\epsilon$	$\cdot$	$\cdot$	$\cdot$
$\mathbf{1}\mathbf{1}^T$	$\frac{1}{n}\delta^2$	$\cdot$	$\frac{2}{n}\epsilon\delta$	$\cdot$	$\cdot$	$\frac{1}{n}\epsilon^2$	$\cdot$
D	$\cdot$	$d_{\text{av}}-2d_{\text{av},2}\epsilon$	1	$-\frac{2}{n}-\frac{2}{n-1}d_{\text{av}}\epsilon$	$\frac{2}{n}\epsilon$	$-(2+\frac{2}{n})\epsilon$	$-\frac{2}{n}\frac{2n-}{n-}$
$d\mathbf{1}^T$	$O(\epsilon^2)$	$\frac{2}{n-1}d_{\text{av}}\delta$	$\cdot$	$\frac{2}{n-1}(d_{\text{av}}\epsilon-\delta)$	$-\frac{2}{n}\epsilon$	$\cdot$	$-\frac{2}{n(n-}$
dd^T	$\cdot$	$d_{\text{av},2}$	$\cdot$	$-2d_{\text{av}}$	1	$\cdot$	$\cdot$
D^2	$\cdot$	$d_{\text{av},2}$	$\cdot$	$\cdot$	$\cdot$	1	$-\frac{2}{n}$
$D^2\mathbf{1}\mathbf{1}^T$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$

Each term "$\cdot$" eventually contributes $O(\epsilon^3)$ or higher to the expected function (though individually these terms can be $O(\epsilon)$, $O(\epsilon^2)$, or $O(\epsilon^3)$).

The terms D, dd^T, D^2, and I are preserved, modulo the decrease by $(1-\delta)^2$.

Expected Attenuation over One Epoch

All scaled by $(1-\delta)^2$.

	I	$\mathbf{1}\mathbf{1}^T$	D	$d\mathbf{1}^T$	dd^T	D^2	$D^2\mathbf{1}\mathbf{1}^T$
I	$1-\frac{2}{n}\delta$	$1-\frac{2}{n}-\frac{2n}{n-1}\epsilon d_{\text{av}}$	$2\frac{n+1}{n}\epsilon$	$\frac{3n-2}{n(n-1)}\epsilon$	$\cdot$	$\cdot$	$\cdot$
$\mathbf{1}\mathbf{1}^T$	$\frac{1}{n}\delta^2$	$\cdot$	$\frac{2}{n}\epsilon\delta$	$\cdot$	$\cdot$	$\frac{1}{n}\epsilon^2$	$\cdot$
D	$\cdot$	$d_{\text{av}}-2d_{\text{av},2}\epsilon$	1	$-\frac{2}{n}-\frac{2}{n-1}d_{\text{av}}\epsilon$	$\frac{2}{n}\epsilon$	$-(2+\frac{2}{n})\epsilon$	$-\frac{2}{n}\frac{2n-1}{n-1}\epsilon$
$d\mathbf{1}^T$	$O(\epsilon^2)$	$\frac{2}{n-1}d_{\text{av}}\delta$	$\cdot$	$\frac{2}{n-1}(d_{\text{av}}\epsilon-\delta)$	$-\frac{2}{n}\epsilon$	$\cdot$	$-\frac{2}{n(n-1)}\epsilon$
dd^T	$\cdot$	$d_{\text{av},2}$	$\cdot$	$-2d_{\text{av}}$	1	$\cdot$	$\cdot$
D^2	$\cdot$	$d_{\text{av},2}$	$\cdot$	$\cdot$	$\cdot$	1	$-\frac{2}{n}$
$D^2\mathbf{1}\mathbf{1}^T$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$

Each term "$\cdot$" eventually contributes $O(\epsilon^3)$ or higher to the expected function (though individually these terms can be $O(\epsilon)$, $O(\epsilon^2)$, or $O(\epsilon^3)$).

The terms D, dd^T, D^2, and I are preserved, modulo the decrease by $(1-\delta)^2$.

The First Few Epochs

Approximately:

	0	1	2
I	δ	δ	δ
$\mathbf{1}\mathbf{1}^T$	$1-\delta$	$\delta+\epsilon d_{\text{av}}$	$\delta+\epsilon d_{\text{av}}$
D	ϵ	ϵ	ϵ
$d\mathbf{1}^T+\mathbf{1}d^T$	0	$-\frac{2}{n}\epsilon$	$-\frac{2}{n}\epsilon$
dd^T	0	$\frac{2}{n}\epsilon^2$	$\frac{2}{n}\epsilon^2$
D^2	0	$-2\epsilon^2$	$-2\epsilon^2$
$D^2\mathbf{1}\mathbf{1}^T+\mathbf{1}\mathbf{1}^TD^2$	0	$-\frac{4}{n}\epsilon^2$	$-\frac{4}{n}\epsilon^2$

The coefficients stabilize after the first epoch!

Recall that there is an additional factor of $(1-\delta)^2$ per epoch. This overriding rate is modified when higher-order effects in the coefficients are accounted for, but the overall trend of a constant decrease rate after the first epoch is still observed.

First Iteration

As for the invariant matrix, f decreases a lot on the first *iteration* for all CD variants, when an exact line search is used.

Lemma

After a single iteration, we have

$$\mathbb{E}_{x^0}\mathbb{E}_i f(x^1) \leq \frac{n-1}{2}\left[\delta + \epsilon d_{av} + \frac{\left(\delta(1-\delta)^2 + (1-\delta)(\delta+\epsilon)^2 + (1-\delta)^2\epsilon\right)}{(1+\epsilon)^2}\right] \tag{10}$$

where i is the coordinate chosen for updating in the first iteration. When $\delta = O(\epsilon)$ and $\epsilon \ll 1$, we have

$$\mathbb{E}_{x^0}\mathbb{E}_i f(x^1) \leq \frac{1}{2}(n-1)[2\delta + \epsilon(d_{av}+1)] + O(\epsilon^2). \tag{11}$$

Same estimates hold for CCD (without the expections with respect to i).

Short version: $f(x^0) \sim n(1+\epsilon)$ while $f(x^1) \sim n(\delta+\epsilon)$.

Summary

Performance bounds for CCD and RPCD are much worse than for RCD, yet performance is similar on many problems.

But there are problems for which the worst-case bounds for CCD are realized, and RCD is much faster.

We prove that on these problems, Random-Permutations Cyclic (RPCD) behaves like fully-random CD — perhaps even a little better.

(For Stochastic Gradient, analogous results comparing Random-Permutations with Fully Random are scarce.)

References I

Beck, A. and Tetruashvili, L. (2013).
On the convergence of block coordinate descent type methods.
SIAM Journal on Optimization, 23(4):2037–2060.

Griewank, A. and Walther, A. (2008).
Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation.
Frontiers in Applied Mathematics. SIAM, Philadelphia, PA, second edition.

Lee, C.-p. and Wright, S. J. (2016).
Random permutations fix a worst case for cyclic coordinate descent.
Technical Report arXiv:1607.08320, Computer Sciences Department, University of Wisconsin-Madison.

Lee, C.-p. and Wright, S. J. (2017).
Analysis of good cases for random-permutations cyclic coordinate descent.
Technical report, Computer Sciences Department, University of Wisconsin-Madison.
In preparation.

Nesterov, Y. (2012).
Efficiency of coordinate descent methods on huge-scale optimization problems.
SIAM Journal on Optimization, 22:341–362.

References II

Sun, R. and Ye, Y. (2016).

Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version.
Technical Report arXiv:1604.07130, Department of Management Science and Engineering, Stanford University, Stanford, California.