Provable Benefit of Random Permutations over Uniform Sampling in Stochastic Coordinate Descent

Donghwa Kim Jaewook Lee Chulhee Yun

Kim Jaechul Graduate School of AI, KAIST

ICML 2025

KAIST AI
Kim Jaechul Graduate School

OptiML Optimization & Machine Learning Laboratory

- **Coordinate Descent (CD)**:
  - Updates one coordinate at a time.
  - Efficient for high-dimensional / large-scale problems.
- **Random Coordinate Descent (RCD)**:
  - Selects a coordinate randomly **with replacement** at each iteration.
- **Random-Permutation Coordinate Descent (RPCD)**:
  - Generates a random permutation **without replacement** at each epoch.

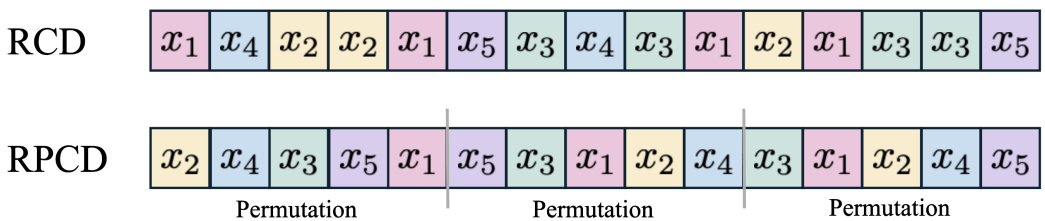

Figure: Illustration of the difference between RCD and RPCD update orders.

- **Empirical Observations**:
  - RPCD often outperforms RCD.
- **Theoretical Gap**:
  - Lack of rigorous proof for RPCD's superiority over RCD.
- **Fundamental Question**:
  - Can we provide a theoretical explanation for RPCD's faster convergence?

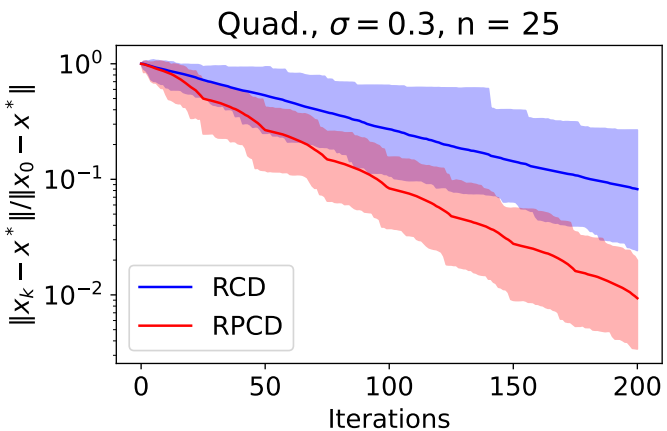

Figure: Empirical performance comparison of RCD and RPCD on a quadratic function.

- **Problem**: Unconstrained quadratic minimization

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x},$$

  where the Hessian $\boldsymbol{A} \in \mathbb{S}_+^n$ is a *positive definite* matrix.
- We define $\sigma := \lambda_{\min}(\boldsymbol{D}^{-1}\boldsymbol{A})$ ($\boldsymbol{D}$ is the diagonal part of $\boldsymbol{A}$).
- Without loss of generality, we assume that $\boldsymbol{A}$ is unit-diagonal (*i.e.*, $a_{ii} = 1$ for all i). In this case, $\boldsymbol{D} = \boldsymbol{I}$, so $\sigma = \lambda_{\min}(\boldsymbol{A})$.

Theorem 3.1 (RCD Lower Bound)

For an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\boldsymbol{x}_T$ be the output of *RCD* after T iterates. Then, except for a Lebesgue measure zero set of initial points,

$$\lim_{T\to\infty} \left(\frac{\mathbb{E}\left[\|\boldsymbol{x}_T\|^2\right]}{\|\boldsymbol{x}_0\|^2} \right)^{\frac{1}{T}} \geq \max\left\{ \left(1 - \frac{1}{n}\right), \left(1 - \frac{\sigma}{n}\right)^2 \right\}.$$

- This theorem establishes a lower bound on the convergence rate of RCD for all quadratic functions with positive definite Hessian.

Definition

We define the class of Hessians $\mathcal{A}_\sigma^{\mathsf{PI}}, \mathcal{A}_\sigma$ as

$$\mathcal{A}_\sigma^{\mathsf{PI}} := \left\{ \mathrm{diag}\{\sigma \boldsymbol{I}_k + (1-\sigma)\mathbf{1}_k \mathbf{1}_k^\top, \boldsymbol{I}_{n-k}\} : 2 \le k \le n \right\},$$

$$\mathcal{A}_\sigma := \left\{ \boldsymbol{A} = \boldsymbol{A}^{\mathsf{PI}} \odot \boldsymbol{v}\boldsymbol{v}^\top : \boldsymbol{A}^{\mathsf{PI}} \in \mathcal{A}_\sigma^{\mathsf{PI}}, \boldsymbol{v} \in \{\pm 1\}^n \right\}$$

for given $\sigma \in (0, 1]$.

- **Note**: If $\boldsymbol{A} \in \mathcal{A}_\sigma$, $\sigma = \lambda_{\min}(\boldsymbol{A})$.
- **Example**:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0.7 & -0.7 & 0 \\ 0.7 & 1 & -0.7 & 0 \\ -0.7 & -0.7 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (\sigma = 0.3).$$

Theorem 3.3 (RPCD Upper Bound)

For an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\boldsymbol{x}_K$ be the output of *RPCD* after K epochs. If $\boldsymbol{A} \in \mathcal{A}_\sigma$ and $\boldsymbol{x}_0 \neq 0$,

$$\lim_{K\to\infty} \left(\frac{\mathbb{E}\left[\|\boldsymbol{x}_K\|^2\right]}{\|\boldsymbol{x}_0\|^2} \right)^{\frac{1}{K}} \leq \max\left\{ \left(1 - \frac{1}{n}\right)^n, \left(1 - \frac{\sigma}{n}\right)^{2n} \right\}.$$

- This upper bound matches the lower bound for RCD in **Theorem 3.1**.

Theorem 3.4 (Stronger RCD Lower Bound)

Let $\boldsymbol{A} \in \mathcal{A}_\sigma$. For an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\boldsymbol{x}_T$ be the output of *RCD* after T iterates. Then, except for a Lebesgue measure zero set of initial points,

$$\lim_{T\to\infty} \left(\frac{\mathbb{E}\left[\|\boldsymbol{x}_T\|^2\right]}{\|\boldsymbol{x}_0\|^2} \right)^{\frac{1}{T}} \geq 1 - \frac{1}{n} + \frac{(1-\sigma)^2}{n}.$$

- This theorem provides a *tighter lower bound* for RCD specifically in the class $\mathcal{A}_\sigma$.
- This bound is *strictly larger* than RPCD's upper bound from **Theorem 3.3**.

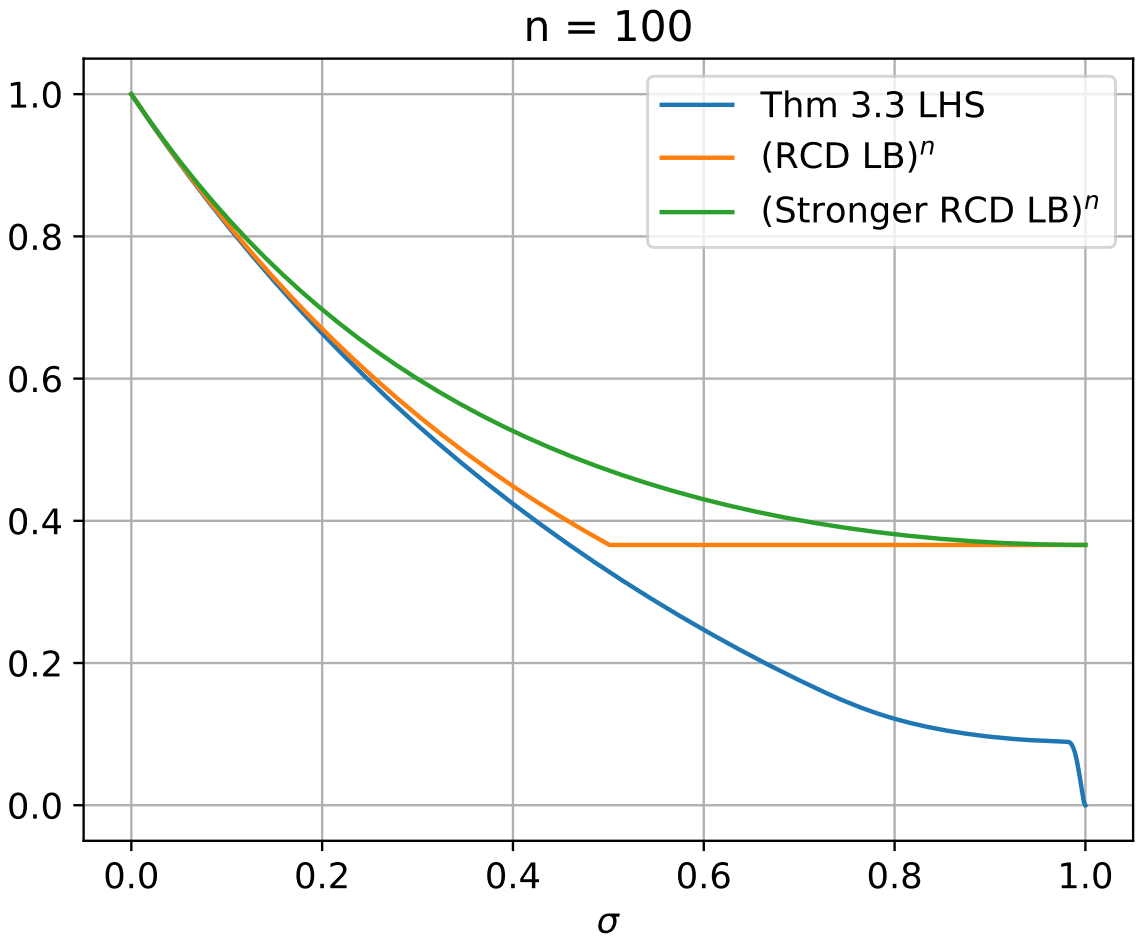

Figure: Theoretical convergence bounds of RCD and RPCD.

Conjecture 4.1

For an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\boldsymbol{x}_K$ be the output of *RPCD* after K epochs. If $\sigma \in (0, 1]$, $\boldsymbol{A} \in \mathbb{S}^n_+$ with $\lambda_{\min}(\boldsymbol{A}) = \sigma$, and $\boldsymbol{x}_0 \neq 0$, then

$$\lim_{K \to \infty} \left(\frac{\mathbb{E}\left[\|\boldsymbol{x}_K\|^2 \right]}{\|\boldsymbol{x}_0\|^2} \right)^{\frac{1}{K}} \leq \max \left\{ \left(1 - \frac{1}{n}\right)^n, \left(1 - \frac{\sigma}{n}\right)^{2n} \right\}.$$

- We conjecture that RPCD is faster than RCD for *all* quadratic functions with positive definite Hessian.

Conclusion
RCD vs RPCD

- **Main Result**: RPCD achieves a strictly better contraction rate than RCD on quadratics with permutation-invariant structures.
- **Significance**: First theoretical result showing RPCD's advantage over RCD.
- **Open Question**: Does this advantage extend to *all* positive definite matrices?