Randomized iterative methods: acceleration and applications

谢家新

与韩德仁教授，戚厚铎教授，苏宴生和曾韵合作完成

xiejx@buaa.edu.cn

北京航空航天大学
数学科学学院

暨南大学
2024 年 12 月 1 日

Contents

1. Randomized Kaczmarz for linear systems
2. Randomized Douglas-Rachford method
3. Adaptive stochastic heavy ball momentum
4. Randomized iterative methods for GAVE

Linear systems

- We consider solving large-scale system of linear equations

$$Ax = b, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

Linear systems

- We consider solving large-scale system of linear equations

$$Ax = b, \tag{1.1}$$

  where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

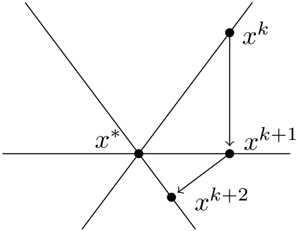
- We assume that this linear system is **consistent**, i.e. there exists a vector x^* for which $Ax^* = b$.

Randomized Kaczmarz method

Kaczmarz method

The Kaczmarz method (1937) iterates as follows:

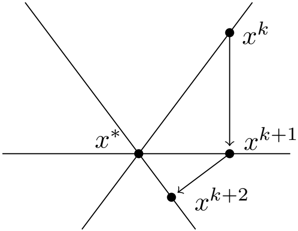
$$x^{k+1} = P_{C_i}(x^k) = x^k - \frac{\langle a_i, x^k \rangle - b_i}{\|a_i\|_2^2} a_i.$$



Kaczmarz method

The Kaczmarz method (1937) iterates as follows:

$$x^{k+1} = P_{C_i}(x^k) = x^k - \frac{\langle a_i, x^k \rangle - b_i}{\|a_i\|_2^2} a_i.$$



By choosing the index i **cyclically**, the iteration sequence $\{x^k\}$ converges to a certain solution x^* of $Ax = b$. The rate of convergence is **hard** to obtain.

RK for consistent linear systems

Strohmer and Vershynin [1] first analyzed the **randomized** variant of the Kaczmarz method. They chose the i-th hyperplane C_i with probability

$$\mathrm{Pr}(\text{row} = i) = \frac{\|a_i\|_2^2}{\|A\|_F^2}. \tag{1.2}$$

[1] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl., 15(2):262–278, 2009.

RK for consistent linear systems

Strohmer and Vershynin [1] first analyzed the **randomized** variant of the Kaczmarz method. They chose the i-th hyperplane C_i with probability

$$\mathrm{Pr}(\mathrm{row} = i) = \frac{\|a_i\|_2^2}{\|A\|_F^2}. \tag{1.2}$$

The convergence result:

$$\mathbb{E}\left[\left\|x^k - x_*^0\right\|_2^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^k \|x^0 - x_*^0\|_2^2, \tag{1.3}$$

where $x_*^0 = A^\dagger b + (I - A^\dagger A)x^0$,

[1] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl., 15(2):262–278, 2009.

RK for consistent linear systems

Strohmer and Vershynin [1] first analyzed the **randomized** variant of the Kaczmarz method. They chose the i-th hyperplane C_i with probability

$$\Pr(\text{row} = i) = \frac{\|a_i\|_2^2}{\|A\|_F^2}. \tag{1.2}$$

The convergence result:

$$\mathbb{E}\left[\left\|x^k - x_*^0\right\|_2^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^k \|x^0 - x_*^0\|_2^2, \tag{1.3}$$

where $x_*^0 = A^\dagger b + (I - A^\dagger A)x^0$, i.e. RK converges with iteration complexity

$$O\left(\log(1/\varepsilon)\|A\|_F^2/\sigma_{\min}^2(A)\right).$$

[1] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl., 15(2):262–278, 2009.

Subsequent work

- Block randomized Kaczmarz [Needell-Tropp, LAA, 2014],

$$x^{k+1} = x^k - (A_{\mathcal{I}_k})^\dagger (A_{\mathcal{I}_k} x^k - b_{\mathcal{I}_k}).$$

- Accelerated randomized Kaczmarz [Liu-Wright, Math. Comp., 2015],

$$y^{k+1} = x^k + \beta_k (x^k - x^{k-1}),$$
$$x^{k+1} = y^k - \frac{\langle a_{i_k}, y^k \rangle - b_{i_k}}{\|a_{i_k}\|_2^2} a_{i_k}.$$

- Sketch-and-project [Gower-Richtárik, SIMAX, 2015],

$$\min \|x - x^k\|_B^2 \ \text{ s.t. } \ S^\top A x = S^\top b,$$

i.e.

$$x^{k+1} = x^{k+1} - B^{-1} A^\top S (S^\top A B^{-1} A S^\top)^\dagger S^\top (A x^k - b).$$

Subsequent work

- Randomized sparse Kaczmarz [Schöpfer-Lorenz, MP, 2019],

$$\begin{aligned} z^{k+1} &= z^k - \frac{\langle a_{i_k}, x^k\rangle - b_{i_k}}{\|a_{i_k}\|_2^2} a_{i_k}, \\ x^{k+1} &= \mathcal{S}_\lambda(z^{k+1}), \end{aligned}$$

  where $\mathcal{S}$ is the soft shrinkage operator.

- Randomized average block Kaczmarz [Necoara, SIMAX, 2019],

$$x^{k+1} = x^k - \alpha_k \left(\sum_{i \in \mathcal{J}_k} \omega_i^k \frac{\langle a_i, x^k\rangle - b_i}{\|a_i\|_2^2} a_i^\top \right),$$

  where the weights $\omega_i^k \in [0,1]$ such that $\sum\limits_{i \in \mathcal{J}_k} \omega_i^k = 1$, $\mathcal{J}_k \subseteq [m]$, and $\alpha_k > 0$ is the step-size.

Subsequent work

- Tensor recovery [Chen-Qin, SIIMS, 2021], Bregman-Kaczmarz method [Yuan-Zhang-Wang-Zhang, IP, 2022], consider the following optimization problem

  $$\min_x f(x) \ \text{ s.t. } \ Ax = b.$$

  The Bregman-Kaczmarz has the following iteration scheme

  $$\begin{aligned} z^{k+1} &= z^k - \frac{\langle a_{i_k}, x^k \rangle - b_{i_k}}{\|a_{i_k}\|_2^2} a_{i_k}, \\ x^{k+1} &= \nabla f^*(z^{k+1}). \end{aligned}$$

Subsequent work

- Tensor recovery [Chen-Qin, SIIMS, 2021], Bregman-Kaczmarz method [Yuan-Zhang-Wang-Zhang, IP, 2022], consider the following optimization problem

  $$\min_x f(x) \ \text{ s.t. } \ Ax = b.$$

  The Bregman-Kaczmarz has the following iteration scheme

  $$\begin{aligned} z^{k+1} &= z^k - \frac{\langle a_{i_k}, x^k\rangle - b_{i_k}}{\|a_{i_k}\|_2^2} a_{i_k}, \\ x^{k+1} &= \nabla f^*(z^{k+1}). \end{aligned}$$

- Randomized Douglas-Rachford [Han-Su-Xie, SIOPT, 2024]
- Adaptive stochastic heavy ball momentum [Zeng-Han-Su-Xie, SIMAX, 2024]
- ......

Douglas-Rachford method

Douglas-Rachford method

The Douglas-Rachford (DR) method is a very popular method for finding zeroes of sums of **maximally monotone operators**.

- splitting algorithms
- ADMM (alternating direction method of multipliers)
- convex optimization

Douglas-Rachford method

The Douglas-Rachford (DR) method is a very popular method for finding zeroes of sums of **maximally monotone operators**.

- splitting algorithms
- ADMM (alternating direction method of multipliers)
- convex optimization
- nonconvex feasibility problems
- ......

The original DR method is restricted to solving

$$\text{Find } x^* \in C_1 \cap C_2.$$

The original DR method is restricted to solving

$$\text{Find } x^* \in C_1 \cap C_2.$$

Its iterative scheme reads as

$$x^{k+1} = \frac{1}{2}(I + R_{C_2} R_{C_1})(x^k), \quad k \geq 0,$$

where $R_{C_i} := 2P_{C_i} - I$ denotes the **reflection** through the set C_i.

The original DR method is restricted to solving

$$\text{Find } x^* \in C_1 \cap C_2.$$

Its iterative scheme reads as

$$x^{k+1} = \frac{1}{2}(I + R_{C_2} R_{C_1})(x^k), \quad k \geq 0,$$

where $R_{C_i} := 2P_{C_i} - I$ denotes the **reflection** through the set C_i.

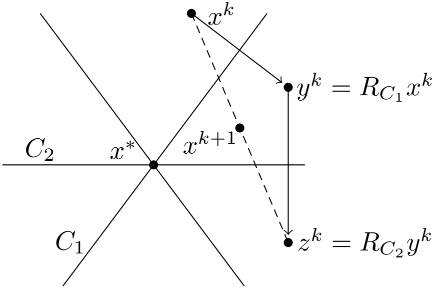


Figure: One step of the Douglas-Rachford method: **Reflect-reflect-average**; $x^{k+1} = \frac{1}{2}(x^k + z^k) = \frac{1}{2}(I + R_{C_2} R_{C_1}) x^k$.

Direct extension of DR may fail

Find $x^* \in \bigcap_{i=1}^{3} C_i$, the direct extension of 3-sets-DR

$$x^{k+1} = \frac{1}{2}(I + R_{C_3} R_{C_2} R_{C_1})(x^k)$$

may **fail**.

Direct extension of DR may fail

Find $x^* \in \bigcap_{i=1}^{3} C_i$, the direct extension of 3-sets-DR

$$x^{k+1} = \frac{1}{2}(I + R_{C_3}R_{C_2}R_{C_1})(x^k)$$

may **fail**.

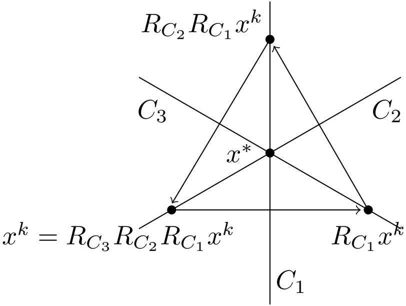


Figure: Failure of the 3-sets-Douglas-Rachford iteration: The iteration $x^k := \left(\frac{1}{2}I + \frac{1}{2}R_{C_3}R_{C_2}R_{C_1}\right)^k x^0$ may **cycle**.

The cyclic-DR algorithm

For an arbitrary x_0, the cyclic-DR employed the sequential iterative process

$$x^{k+1} = \frac{1}{2}(I + R_{C_{t_k}} R_{C_{t_{k-1}}})(x^k) \tag{2.1}$$

where $t_k = k \bmod m + 1$.

The cyclic-DR algorithm

For an arbitrary x_0, the cyclic-DR employed the sequential iterative process

$$x^{k+1} = \frac{1}{2}(I + R_{C_{t_k}} R_{C_{t_{k-1}}})(x^k) \tag{2.1}$$

where $t_k = k \bmod m + 1$. One can general this idea to obtain the **cyclic r-sets-DR method** (r reflections at each step).

The cyclic-DR algorithm

For an arbitrary x_0, the cyclic-DR employed the sequential iterative process

$$x^{k+1} = \frac{1}{2}(I + R_{C_{t_k}} R_{C_{t_{k-1}}})(x^k) \tag{2.1}$$

where $t_k = k \bmod m + 1$. One can general this idea to obtain the **cyclic r-sets-DR method** (r reflections at each step).

- Theoretical estimates of the rate of convergence of the (cyclic)-DR method are difficult $(m > 2)$;
- Known estimates for the rate of convergence are hard to compute;

Motivation

Is it possible that a carefully designed randomization scheme makes the (otherwise divergent) r-set-DR method convergent?

Randomized Douglas-Rachford method

The consideration

We consider the direct extension of the extrapolated r**-sets-DR method** with the following iterative procedure

$$x^{k+1} = \left((1-\alpha)I + \alpha R_{C_{j_{k_r}}} R_{C_{j_{k_{r-1}}}} \ldots R_{C_{j_{k_1}}}\right)(x^k),$$

where $\alpha \in (0,1)$ is the extrapolation or relaxation parameter.

The consideration

We consider the direct extension of the extrapolated **r-sets-DR method** with the following iterative procedure

$$x^{k+1} = \left((1-\alpha)I + \alpha R_{C_{j_{k_r}}} R_{C_{j_{k_{r-1}}}} \dots R_{C_{j_{k_1}}}\right)(x^k),$$

where $\alpha \in (0,1)$ is the extrapolation or relaxation parameter.

- If we take $r = 1$ and $\alpha = \frac{1}{2}$, it recovers the Kaczmarz method.
- If we take $r = 2$ and $\alpha = \frac{1}{2}$, it recovers the classical DR method.

Randomized r-sets-Douglas-Rachford method

```
Algorithm 1: Randomized $r$-sets-Douglas-Rachford (RrDR) algorithm .
Input: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $0 < r \in \mathbb{Z}_+$, $k = 0$, extrapolation/relaxation parameter $\alpha \in (0,1)$ and an initial $x_0 \in \mathbb{R}^n$.
Step 1. Set $z_0^k := x^k$.
Step 2. for $\ell = 1, \ldots, r$ do
            Select $j_{k_\ell} \in \{1, \ldots, m\}$ with probability $\Pr(\text{row} = j_{k_\ell}) = \frac{\|a_{j_{k_\ell}}\|_2^2}{\|A\|_F^2}$.
            Compute $z_\ell^k := z_{\ell-1}^k - 2\frac{\langle a_{j_{k_\ell}}, z_{\ell-1}^k\rangle - b_{j_{k_\ell}}}{\|a_{j_{k_\ell}}\|_2^2} a_{j_{k_\ell}}$.
        end for
Step 3. Update
            $$x^{k+1} := (1-\alpha)x^k + \alpha z_r^k.$$
Step 4. If the stopping rule is satisfied, stop and go to output. Otherwise, set
        $k = k + 1$ and return to Step 1.
Output: The approximate solution.
```

Convergence of iterates: Linear rate

Algorithm 1 converges with iteration complexity

$$O\left(\left(\alpha\left(1-\left(1-2\frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^r\right)\right)^{-1}\log(1/\varepsilon)\right)[2],$$

[2] D. Han, Y. Su, and J. Xie. Randomized Douglas-Rachford method for linear systems: Improved accuracy and efficiency, SIAM J. Optim. 34(1) 1045–1070, 2024.

Heavy ball momentum acceleration

Gradient descent

Consider the following unconstrained minimization problem:

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

where f is a differentiable convex function. Gradient descent (GD) uses the following update formula

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right).$$

Gradient descent

Consider the following unconstrained minimization problem:

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

where f is a differentiable convex function. Gradient descent (GD) uses the following update formula

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right).$$

Iteration complexity:

$$O((L/\mu)\log(1/\varepsilon))$$

with f being L-smooth and μ-strongly convex.

The heavy ball method

Polyak proposed the following momentum variant of the gradient descent method

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right) + \beta_{\mathrm{k}}\left(\mathrm{x}^{\mathrm{k}} - \mathrm{x}^{\mathrm{k}-1}\right).$$

The heavy ball method

Polyak proposed the following momentum variant of the gradient descent method

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right) + \beta_{\mathrm{k}}\left(\mathrm{x}^{\mathrm{k}} - \mathrm{x}^{\mathrm{k}-1}\right).$$

Iteration complexity:

$$O(\sqrt{L/\mu}\log(1/\varepsilon)),$$

with α_k and β_k being properly chosen.

The heavy ball method

Polyak proposed the following momentum variant of the gradient descent method

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right) + \beta_{\mathrm{k}}\left(\mathrm{x}^{\mathrm{k}} - \mathrm{x}^{\mathrm{k}-1}\right).$$

Iteration complexity:

$$O(\sqrt{L/\mu}\log(1/\varepsilon)),$$

with α_k and β_k being properly chosen.
Quadratic improvement.

The heavy ball method

Polyak proposed the following momentum variant of the gradient descent method

$$x^{k+1} = x^k - \alpha_k \nabla f\left(x^k\right) + \beta_{\mathrm{k}}\left(\mathrm{x}^{\mathrm{k}} - \mathrm{x}^{\mathrm{k}-1}\right).$$

Iteration complexity:

$$O(\sqrt{L/\mu}\log(1/\varepsilon)),$$

with α_k and β_k being properly chosen.
Quadratic improvement.

Is it possible for the DR method to achieve accelerated convergence rates with heavy ball momentum?

Algorithm 2: Randomized r-sets-Douglas-Rachford with momentum (mRrDR).

```
Input: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $0 < r \in \mathbb{Z}_+$, $k = 1$, relaxation parameter $\alpha$, the heavy ball/momentum parameter $\beta$, and initial vectors $x^1, x^0 \in \mathbb{R}^n$.
Step 1. Set $z_0^k := x^k$.
Step 2. for $\ell = 1, \ldots, r$ do
            Select $j_{k_\ell} \in \{1, \ldots, m\}$ with probability $\Pr(\text{row} = j_{k_\ell}) = \frac{\|a_{j_{k_\ell}}\|_2^2}{\|A\|_F^2}$.
            Compute $z_\ell^k := z_{\ell-1}^k - 2\frac{\langle a_{j_{k_\ell}}, z_{\ell-1}^k\rangle - b_{j_{k_\ell}}}{\|a_{j_{k_\ell}}\|_2^2} a_{j_{k_\ell}}$.
         end for
Step 3. Update
            $x^{k+1} := (1-\alpha)x^k + \alpha z_r^k + \beta(x^k - x^{k-1})$.
Step 4. If the stopping rule is satisfied, stop and go to output. Otherwise, set
        $k = k + 1$ and return to Step 1.
Output: The approximate solution.
```

Accelerated linear rate

Algorithm 2 converges with iteration complexity:

$$O\left(\sqrt{\left(0.99\alpha\left(1-\left(1-2\frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^r\right)\right)^{-1}}\log(1/\varepsilon)\right)$$ [3],

provided the momentum parameter

$$\beta=\left(1-\sqrt{0.99\alpha\left(1-\left(1-2\frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^r\right)}\right)^2.$$

Quadratic improvement.

[3] D. Han, Y. Su, and J. Xie. Randomized Douglas-Rachford method for linear systems: Improved accuracy and efficiency. SIAM J. Optim. 34(1) 1045–1070, 2024.

Comparison to the cyclic DR method

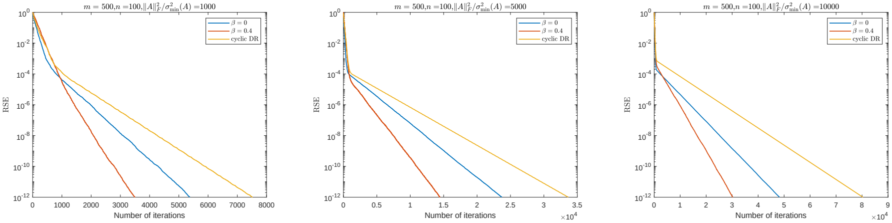


Figure: Comparison of mRrDR $(r = 2, \alpha = 0.5)$ and the cyclic DR $(\alpha = 0.5)$.

Adaptive stochastic heavy ball momentum

Stochastic heavy ball momentum method

- The stochastic gradient descent (SGD) for solving the finite-sum problem $\min_x \frac{1}{m}\sum_{i=1}^m f_i(x)$ utilizes the update

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

  where α_k is the step-size and i_k is selected randomly.
- The stochastic heavy ball momentum (SHBM) method utilizes

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

A limitation of SHBM

- The optimal choices of α_k and β_k for HBM for solving $\min_{x\in\mathbb{R}^n}\frac{1}{2}\|Ax-b\|_2^2$ are [4]

$$\sqrt{\alpha^*}=\frac{2}{\sigma_{\max}(A)+\sigma_{\min}(A)} \text{ and } \sqrt{\beta^*}=\frac{\alpha^*\left(\sigma_{\max}^2(A)-\sigma_{\min}^2(A)\right)}{4},$$

- SHBM for solving the linear system $Ax=b$, i.e. RK with momentum, also require the knowledge of singular values of A [5] [6].
- Is it possible to design a theoretically supported adaptive SHBM method? [7]

[4] Boris T Polyak. Comput. Math. Math. Phys., 4(5):1–17, 1964.

[5] Nicolas Loizou and Peter Richtárik. Comput. Optim. Appl., 77(3):653–710, 2020.

[6] Deren Han and Jiaxin Xie. arXiv: 2208.05437, 2022

[7] Mathieu Barré, Adrien Taylor, and Alexandre d'Aspremont. In Conference on Learning Theory, pages 452¨C478. PMLR, 2020.

SHBM for consistent linear systems

- Consider the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) := \mathbb{E}\left[f_S(x)\right], \tag{3.1}$$

  with

$$f_S(x) := \frac{1}{2}\left\|S^\top(Ax-b)\right\|_2^2 .$$

- At the k-th iteration $(k \geq 1)$, the SHBM method for solving linear systems updates

$$x^{k+1} = x^k - \alpha_k \nabla f_{S_k}(x^k) + \beta_k(x^k - x^{k-1}),$$

  where $\nabla f_{S_k}(x^k) = A^\top S_k S_k^\top(Ax^k - b)$ and S_k is randomly chosen from the probability space $(\Omega_k, \mathcal{F}_k, P_k)$.

SHBM for consistent linear systems

- Consider the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) := \mathbb{E}\left[f_S(x)\right], \tag{3.1}$$

with

$$f_S(x) := \frac{1}{2}\left\|S^\top(Ax-b)\right\|_2^2 .$$

- At the k-th iteration $(k \geq 1)$, the SHBM method for solving linear systems updates

$$x^{k+1} = x^k - \alpha_k \nabla f_{S_k}(x^k) + \beta_k(x^k - x^{k-1}),$$

where $\nabla f_{S_k}(x^k) = A^\top S_k S_k^\top (Ax^k - b)$ and S_k is randomly chosen from the probability space $(\Omega_k, \mathcal{F}_k, P_k)$.

- We intend to choose α_k and β_k to be the minimizers of the following constrained optimization problem

$$\begin{aligned} \min_{\alpha,\beta\in\mathbb{R}} \quad & \|x - A^\dagger b\|_2^2 \\ \text{subject to} \quad & x = x^k - \alpha\nabla f_{S_k}(x^k) + \beta(x^k - x^{k-1}). \end{aligned} \tag{3.2}$$

- If $\|\nabla f_{S_k}(x^k)\|_2^2\|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle^2 \neq 0$, then the minimizers of (3.2) are

$$\begin{cases} \alpha_k = \frac{\|x^k - x^{k-1}\|_2^2\langle \nabla f_{S_k}(x^k), x^k - A^\dagger b\rangle - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle\langle x^k - x^{k-1}, x^k - A^\dagger b\rangle}{\|\nabla f_{S_k}(x^k)\|_2^2\|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle^2}, \\ \beta_k = \frac{\langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle\langle \nabla f_{S_k}(x^k), x^k - A^\dagger b\rangle - \|\nabla f_{S_k}(x^k)\|_2^2\langle x^k - x^{k-1}, x^k - A^\dagger b\rangle}{\|\nabla f_{S_k}(x^k)\|_2^2\|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle^2}, \end{cases} \tag{3.3}$$

- It seems to be intractable since $\langle \nabla f_{S_k}(x^k), x^k - A^\dagger b\rangle$ and $\langle x^k - x^{k-1}, x^k - A^\dagger b\rangle$ include an unknown vector $A^\dagger b$.

- If $\|\nabla f_{S_k}(x^k)\|_2^2\|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1}\rangle^2 \neq 0$, then the minimizers of (3.2) are

$$\begin{cases} \alpha_k = \frac{\|x^k-x^{k-1}\|_2^2\langle\nabla f_{S_k}(x^k),x^k-A^\dagger b\rangle-\langle\nabla f_{S_k}(x^k),x^k-x^{k-1}\rangle\langle x^k-x^{k-1},x^k-A^\dagger b\rangle}{\|\nabla f_{S_k}(x^k)\|_2^2\|x^k-x^{k-1}\|_2^2-\langle\nabla f_{S_k}(x^k),x^k-x^{k-1}\rangle^2}, \\ \beta_k = \frac{\langle\nabla f_{S_k}(x^k),x^k-x^{k-1}\rangle\langle\nabla f_{S_k}(x^k),x^k-A^\dagger b\rangle-\|\nabla f_{S_k}(x^k)\|_2^2\langle x^k-x^{k-1},x^k-A^\dagger b\rangle}{\|\nabla f_{S_k}(x^k)\|_2^2\|x^k-x^{k-1}\|_2^2-\langle\nabla f_{S_k}(x^k),x^k-x^{k-1}\rangle^2}, \end{cases} \tag{3.3}$$

- It seems to be intractable since $\langle \nabla f_{S_k}(x^k), x^k - A^\dagger b\rangle$ and $\langle x^k - x^{k-1}, x^k - A^\dagger b\rangle$ include an unknown vector $A^\dagger b$.
- Noting that $\nabla f_{S_k}(x^k) = A^\top S_k S_k^\top(Ax^k - b)$ and $AA^\dagger b = b$, then we know that

$$\langle \nabla f_{S_k}(x^k), x^k - A^\dagger b\rangle = \langle S_k^\top(Ax^k - b), S_k^\top A(x^k - A^\dagger b)\rangle = \|S_k^\top(Ax^k - b)\|_2^2.$$

Compute $\langle x^k - x^{k-1}, x^k - A^\dagger b \rangle$

- From Figure 3.1, we know that $\langle x^k - x^{k-1}, x^k - A^\dagger b \rangle = 0$.

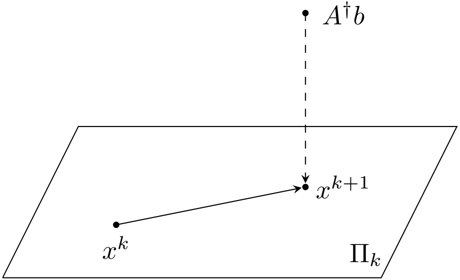


Figure: A geometric interpretation of our design. The next iterate x^{k+1} arises such that x^{k+1} is the orthogonal projection of $A^\dagger b$ onto the affine set $\Pi_k = x^k + \text{Span}\{\nabla f_{S_k}(x^k), x^k - x^{k-1}\}$.

- (3.3) can be simplified to

$$\begin{cases} \alpha_k = \frac{\|x^k - x^{k-1}\|_2^2 \|S_k^\top (Ax^k - b)\|_2^2}{\|\nabla f_{S_k}(x^k)\|_2^2 \|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1} \rangle^2}, \\ \\ \beta_k = \frac{\langle \nabla f_{S_k}(x^k), x^k - x^{k-1} \rangle \|S_k^\top (Ax^k - b)\|_2^2}{\|\nabla f_{S_k}(x^k)\|_2^2 \|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1} \rangle^2}. \end{cases} \tag{3.4}$$

- Overall, for any $k \geq 1$, if

$$\|\nabla f_{S_k}(x^k)\|_2^2 \|x^k - x^{k-1}\|_2^2 - \langle \nabla f_{S_k}(x^k), x^k - x^{k-1} \rangle^2 \neq 0,$$

i.e. $\dim(\Pi_k) = 2$, can be guaranteed, then the minimizer of (3.2) can be computed by (3.4).

A requirement for sampling matrices

Proposition 3.1

Let $x^0 \in \text{Range}(A^\top)$ be an initial point and suppose that x^1 is generated by SGD with stochastic Polyak step-size. Let $\{x^k\}_{k\geq 2}$ be the sequence obtained by solving the optimization problem (3.2). We have the following results:

(i) For any fixed ℓ, if $S_\ell^\top(Ax^\ell - b) \neq 0$, then $x^{\ell+1} \neq x^\ell$;

(ii) If $S_\ell^\top(Ax^\ell - b) \neq 0$ and $S_{\ell-1}^\top(Ax^{\ell-1} - b) \neq 0(\ell \geq 1)$, then $\dim(\Pi_\ell) = 2$.

Proposition 3.1 indicates that if the sampling matrices S_k are chosen such that $S_k^\top(Ax^k - b) \neq 0$ for $k \geq 0$, then $\dim(\Pi_\ell) = 2(\ell \geq 1)$.

ASHBM for linear systems

```
Algorithm 1: Adaptive stochastic heavy ball momentum (ASHBM)
Input: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^{m}$, probability spaces $\{(\Omega_k, \mathcal{F}_k, P_k)\}_{k\geq 0}$, $k = 1$, and the initial point $x^0 \in \text{Range}(A^\top)$.
Step 1. Update $x^1$ by the improved basic method with $\zeta = 1$.
Step 2. Randomly select a sampling matrix $S_k \in \Omega_k$ until $S_k^\top (Ax^k - b) \neq 0$.
Step 3. Compute the parameters $\alpha_k$ and $\beta_k$ in (3.4).
Step 4. Update $x^{k+1} = x^k - \alpha_k A^\top S_k S_k^\top (Ax^k - b) + \beta_k (x^k - x^{k-1})$.
Step 5. If the stopping rule is satisfied, stop and go to output. Otherwise, set
        $k = k + 1$ and return to Step $2$.
Output: The approximate solution.
```

Connection to conjugate gradient-type methods

An equivalent expression for ASHBM

Theorem 3.2 (Zeng-Han-Su-Xie, 2024)

Suppose that $x^0 \in$ *Range*$(A^\top)$ *is the initial point in ASHBM and set* $p_0 = -\nabla f_{S_0}(x^0) = -A^\top S_0 S_0^\top (Ax^0 - b)$. *Then for any* $k \geq 0$, *ASHBM can be equivalently rewritten as*

$$\begin{cases} \delta_k = \|S_k^\top (Ax^k - b)\|_2^2 / \|p_k\|_2^2, \\ x^{k+1} = x^k + \delta_k p_k, \\ \nabla f_{S_{k+1}}(x^{k+1}) = A^\top S_{k+1} S_{k+1}^\top (Ax^{k+1} - b), \\ \eta_k = \langle \nabla f_{S_{k+1}}(x^{k+1}), p_k \rangle / \|p_k\|_2^2, \\ p_{k+1} = -\nabla f_{S_{k+1}}(x^{k+1}) + \eta_k p_k. \end{cases} \tag{3.5}$$

Proposition 3.3

Suppose that $\{x^k\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ are the sequences generated by (3.5). Let $r^k = Ax^k - b$. Then we have

(1) $\langle p_k, p_{k+1}\rangle = 0$;

(2) $(r^{k+1})^\top S_k S_k^\top r^k = 0$.

Conjugate gradient normal equation error(CGNE)

Consider the following equivalent problem

$$AA^\top y = b, x = A^\top y \tag{3.6}$$

of $Ax = b$. Starting with an arbitrary point $x^0 \in \mathbb{R}^n$, $r^0 = Ax^0 - b$, and $p_0 = -A^\top r^0$, the CG method for solving (3.6) results in the following *conjugate gradient normal equation error* (CGNE) method [8]

$$\begin{cases} \mu_k = \frac{\|r^k\|_2^2}{\|p_k\|_2^2}, \\ x^{k+1} = x^k + \mu_k p_k, \\ r^{k+1} = r^k + \mu_k A p_k, \\ \tau_k = \frac{\|r^{k+1}\|_2^2}{\|r^k\|_2^2}, \\ p_{k+1} = -A^\top r^{k+1} + \tau_k p_k. \end{cases} \tag{3.7}$$

[8] Gene H Golub and Charles F Van Loan. Matrix computations. JHU press, 2013

- We have

$$\begin{aligned}\eta_k &= \frac{\langle \nabla f_{S_{k+1}}(x^{k+1}), p_k\rangle}{\|p_k\|_2^2} = \frac{\langle A^\top S_{k+1} S_{k+1}^\top r^{k+1}, p_k\rangle}{\|p_k\|_2^2} \\ &= \frac{\langle S_{k+1}^\top r^{k+1}, S_{k+1}^\top (r^{k+1} - r^k)\rangle}{\delta_k \|p_k\|_2^2} = \frac{\|S_{k+1}^\top r^{k+1}\|_2^2 - \langle S_{k+1}^\top r^{k+1}, S_{k+1}^\top r^k\rangle}{\|S_k^\top r^k\|_2^2}.\end{aligned}$$

- If we assume that the sample spaces $\Omega_k = \{I\}$ for all k, Proposition 3.3 implies that now the parameter $\eta_k = \frac{\|r^{k+1}\|_2^2}{\|r^k\|_2^2}$. It is evident that (3.5) and (3.7) are now equivalent, indicating that CGNE is a special case of ASHBM.

Stochastic conjugate gradient for linear systems

```
Algorithm 3: Stochastic conjugate gradient (SGS)
Input: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, probability spaces $\{(\Omega_k, \mathcal{F}_k, P_k)\}_{k\geq 0}$, $k = 0$ and the initial point $x^0 \in \text{Range}(A^\top)$.
Step 1. Randomly select a sampling matrix $S_0 \in \Omega_0$ until $S_0^\top(Ax^0 - b) \neq 0$.
Step 2. Set $p_0 = -A^\top SS^\top(Ax^0 - b)$.
Step 3. Set $\delta_k = \|S_k^\top(Ax^k - b)\|_2^2 / \|p_k\|_2^2$.
Step 4. Update $x^{k+1} = x^k + \delta_k p_k$.
Step 5. Randomly select a sampling matrix $S_{k+1} \in \Omega_{k+1}$ until
        $S_{k+1}^\top(Ax^{k+1} - b) \neq 0$.
Step 6. Compute
        $$\eta_k = \frac{\|S_{k+1}^\top(Ax^{k+1} - b)\|_2^2 - \langle S_{k+1}^\top(Ax^{k+1} - b), S_{k+1}^\top(Ax^k - b)\rangle}{\|S_k^\top(Ax^k - b)\|_2^2},$$
        $$p_{k+1} = -A^\top S_{k+1} S_{k+1}^\top(Ax^{k+1} - b) + \eta_k p_k.$$
Step 7. If the stopping rule is satisfied, stop and go to output. Otherwise, set
        $k = k + 1$ and return to Step 3.
Output: The approximate solution.
```

Convergence result

Assumption 3.4

Let $\{(\Omega_k, \mathcal{F}_k, P_k)\}_{k\geq 0}$ be a class of probability spaces. We assume that for any $k \geq 0$, $\mathbb{E}_{S\in\Omega_k}\left[SS^\top\right]$ is a positive definite matrix.

Theorem 3.5

Suppose that the linear system $Ax = b$ is consistent and the probability spaces $\{(\Omega_k, \mathcal{F}_k, P_k)\}_{k\geq 0}$ satisfy Assumption 3.4. Let $\{x^k\}_{k\geq 0}$ be the iteration sequence generated by ASHBM. Then

$$\mathbb{E}\left[\|x^{k+1} - A^\dagger b\|_2^2 \mid x^k\right] \leq (1 - \gamma_k)\left(1 - \frac{\sigma_{\min}^2(H_k^{\frac{1}{2}}A)}{\lambda_{\max}^k}\right)\|x^k - A^\dagger b\|_2^2,$$

where γ_k, H_k, and $\lambda_{\max}^k$ are certain parameters. [a]

[a] Zeng-Han-Su-Xie, On adaptive stochastic heavy ball momentum for solving linear systems, SIAM J. Matrix Anal. Appl., 45(3), 1259–1286, 2024.

Comparison to `pinv` and `lsqminnorm`

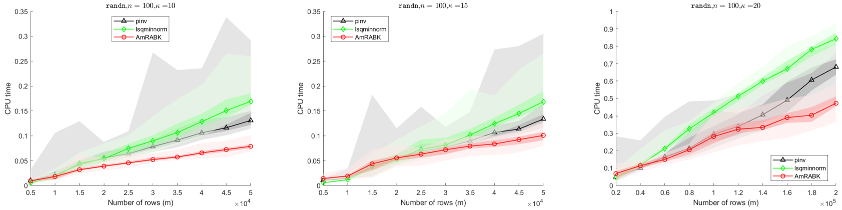


Figure: Figures depict the CPU time (in seconds) vs increasing number of rows. The title of each plot indicates the values of n and κ. We set $p = 30$.

Randomized iterative method for GAVE

Generalized absolute value equations

- We consider the following generalized AVE (GAVE)

$$Ax - B|x| = b, \tag{4.1}$$

where $A, B \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$, $|x| = (|x_1|, \dots, |x_n|)^\top$.

Generalized absolute value equations

- We consider the following generalized AVE (GAVE)

$$Ax - B|x| = b, \tag{4.1}$$

  where $A, B \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^{m}$, $|x| = (|x_1|, \dots, |x_n|)^\top$.
- Many applications, like linear complementarity problem (LCP), biometrics, game theory, etc.

Generalized absolute value equations

- We consider the following generalized AVE (GAVE)
$$Ax - B|x| = b, \tag{4.1}$$
where $A, B \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$, $|x| = (|x_1|, \ldots, |x_n|)^\top$.
- Many applications, like linear complementarity problem (LCP), biometrics, game theory, etc.
- When $m = n$ and B is the identity matrix, the GAVE (4.1) reduces to the standard AVE.
- When $B = 0$, the GAVE (4.1) becomes the system of linear equations.

A fruitful line of research

- Algorithms;
- Solvability;
- Error bound;
- Tensor AVE;
- Applications;
- ......

Algorithms for GAVE

If $m = n$

- Newton methods
- Picard iteration methods
- Splitting iteration methods
- Concave minimization approaches
- ...

Algorithms for GAVE

If $m = n$

- Newton methods
- Picard iteration methods
- Splitting iteration methods
- Concave minimization approaches
- ...

If $m \neq n$

- Successive linearization algorithm via concave minimization
- Method of alternating projections

Picard iteration methods

- Picard iteration method

$$x^{k+1} = A^{-1}(B|x^k| + b).$$

Picard iteration methods

- Picard iteration method

$$x^{k+1} = A^{-1}(B|x^k| + b).$$

- At each step, the following linear system is solved

$$Ax = B|x^k| + b.$$

Picard iteration methods

- Picard iteration method

$$x^{k+1} = A^{-1}(B|x^k| + b).$$

- At each step, the following linear system is solved

$$Ax = B|x^k| + b.$$

- Using a single step of the randomized iterative method to solve the above linear system

$$x^{k+1} = x^k - \alpha \frac{A^\top S_k S_k^\top (Ax^k - B|x^k| - b)}{\|S_k^\top A\|_2^2},$$

where $\alpha > 0$ is the stepsize and $S_k \in \mathbb{R}^{q \times m}$ is a random matrix drawn from a user-defined probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

Randomized iterative method for GAVE

Algorithm 2: Randomized iterative method (RIM) for GAVE

Input: $A, B \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, probability spaces $(\Omega, \mathcal{F}, \mathrm{P})$, $\alpha \in (0,1]$, $k = 0$, and the initial point $x^0 \in \mathbb{R}^n$.

Step 1. Randomly select a sampling matrix $S_k \in \Omega$.

Step 2. Update

$$x^{k+1} = x^k - \alpha\frac{A^\top S_k S_k^\top (Ax^k - B|x^k| - b)}{\|S_k^\top A\|_2^2}.$$

Step 3. If a stopping rule is satisfied, stop and go to output. Otherwise, set $k = k+1$ and return to Step 1.

Output: The approximate solution x^k.

Solvability of GAVE

The unique solvability of GAVE

Theorem 4.1 (Wu-Shen, Optim. Lett., 15: 2017–2024, 2021.)

Suppose that $m = n$. *The GAVE* (4.1) *has a unique solution for any* $b \in \mathbb{R}^n$ *if and only if for any* $D \in [-I_n, I_n]$, *the matrix* $A + BD$ *is nonsingular.*

The unique solvability of GAVE

Theorem 4.1 (Wu-Shen, Optim. Lett., 15: 2017–2024, 2021.)

Suppose that $m = n$. The GAVE (4.1) *has a unique solution for any $b \in \mathbb{R}^n$ if and only if for any $D \in [-I_n, I_n]$, the matrix $A + BD$ is nonsingular.*

Theorem 4.2 (Xie-Qi-Han, 2024)

[a] *The GAVE* (4.1) *has a unique solution for any $b \in \mathbb{R}^m$ if and only if $m = n$ and for any $D \in [-I_n, I_n]$, the matrix $A + BD$ is nonsingular.*

[a] Xie-Qi-Han, Randomized iterative methods for generalized absolute value equations: Solvability and error bounds, **arXiv:2405.04091**, 2024.

1. When $m < n$. For any $A, B \in \mathbb{R}^{m\times n}$, there exists $b \in \mathbb{R}^m$ such that the GAVE (4.1) has infinite solutions.
2. When $m > n$. For any $A, B \in \mathbb{R}^{m\times n}$, there exists $b \in \mathbb{R}^m$ such that the GAVE (4.1) is unsolvable.

Sufficient conditions

Theorem 4.3

Suppose that $m = n$*. If* $\sigma_{\min}(A) > \|B\|_2$*. Then for any* $b \in \mathbb{R}^m$*, the GAVE* (4.1) *has a unique solution.*

Sufficient conditions

Theorem 4.3

Suppose that $m = n$. If $\sigma_{\min}(A) > \|B\|_2$. Then for any $b \in \mathbb{R}^m$, the GAVE (4.1) *has a unique solution.*

Theorem 4.4 (Xie-Qi-Han, 2024)

Suppose that $m \leq n$ and there exists a nonsingular matrix $M \in \mathbb{R}^{m\times m}$ such that $\sigma_m(MA) > \|MB\|_2$. Then for any $b \in \mathbb{R}^m$, the GAVE (4.1) *is solvable.*
In particular, if $m = n$, then the GAVE has a unique solution.

Sufficient conditions

Theorem 4.3

Suppose that $m = n$. If $\sigma_{\min}(A) > \|B\|_2$. Then for any $b \in \mathbb{R}^m$, the GAVE (4.1) *has a unique solution.*

Theorem 4.4 (Xie-Qi-Han, 2024)

Suppose that $m \leq n$ and there exists a nonsingular matrix $M \in \mathbb{R}^{m\times m}$ such that $\sigma_m(MA) > \|MB\|_2$. Then for any $b \in \mathbb{R}^m$, the GAVE (4.1) *is solvable.*
In particular, if $m = n$, then the GAVE has a unique solution.

Theorem 4.5 (Xie-Qi-Han, 2024)

Suppose that $m > n$ and $\mathcal{X}^*$ is non-empty. If there exists a nonsingular matrix $M \in \mathbb{R}^{m\times m}$ such that $\sigma_n(MA) > \|MB\|_2$, then $\mathcal{X}^*$ is singleton.

Error bound

Theorem 4.6 (Zamani-Hladík, Math. Program., 198(1):85–113, 2023)

Assume that $m = n$ *and* $A + D$ *is nonsingular for any* $D \in [-I_n, I_n]$, *then*

$$\|x - x^*\| \leq \max_{D \in [-I_n, I_n]} \| \left(A + D\right)^{-1} \| \cdot \|(Ax - |x| - b)\|, \ \forall x \in \mathbb{R}^n,$$

where $\| \cdot \|$ *represents any vector norm and its induced norm.*

Error bound

Theorem 4.6 (Zamani-Hladík, Math. Program., 198(1):85–113, 2023)

Assume that $m = n$ *and* $A + D$ *is nonsingular for any* $D \in [-I_n, I_n]$, *then*

$$\|x - x^*\| \leq \max_{D \in [-I_n, I_n]} \| (A + D)^{-1} \| \cdot \|(Ax - |x| - b)\|, \ \forall x \in \mathbb{R}^n,$$

where $\| \cdot \|$ *represents any vector norm and its induced norm.*

Theorem 4.7 (Xie-Qi-Han, 2024)

Suppose that $\mathcal{X}^*$ *is non-empty and for any* $D \in [-I_n, I_n]$, *the matrix* $A + BD$ *is full column rank. Then for any* $x^* \in \mathcal{X}^*$ *and nonsingular matrix* $M \in \mathbb{R}^{m \times m}$,

$$\|x - x^*\| \leq \max_{D \in [-I_n, I_n]} \| (MA + MBD)^{\dagger} \| \cdot \|M(Ax - B|x| - b)\|, \ \forall x \in \mathbb{R}^n,$$

where $\| \cdot \|$ *represents any vector norm and its induced norm.*

Convergence result

Theorem 4.8 (Xie-Qi-Han, 2024)

Assume that $\mathcal{X}^*$ *is nonempty and the probability spaces* $(\Omega, \mathcal{F}, \mathrm{P})$ *satisfy Assumption 3.4. Let* $H = \mathbb{E}\left[\frac{SS^\top}{\|S^\top A\|_2^2}\right]$ *and* $\{x^k\}_{k\geq 0}$ *be the iteration sequence generated by Algorithm 2.*

(i) *If* $\sigma_n(H^{\frac{1}{2}}A) = \|H^{\frac{1}{2}}B\|_2$ *and* $\alpha \in (0,1)$, *then at least one subsequence of* $\{x^k\}_{k\geq 0}$ *converges a.s. to a point in the set* $\mathcal{X}^*$ *and* $Ax^k - B|x^k| - b$ *converges a.s. to zero.*

(ii) *If* $\sigma_n(H^{\frac{1}{2}}A) > \|H^{\frac{1}{2}}B\|_2$ *and* $\alpha = (2-\xi)\frac{\sigma_n(H^{\frac{1}{2}}A)}{\sigma_n(H^{\frac{1}{2}}A) - \|H^{\frac{1}{2}}B\|_2}$ *with* $\xi \in \left[\frac{\sigma_n(H^{\frac{1}{2}}A) + \|H^{\frac{1}{2}}B\|_2}{\sigma_n(H^{\frac{1}{2}}A)}, 2\right)$, *then* x^* *is the unique solution and*

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \leq \left(1 - (2-\xi)\xi\sigma_n^2(H^{\frac{1}{2}}A)\right)^k \|x^0 - x^*\|_2^2.$$

Comparison of PIM, GNM, MAP, and RABK

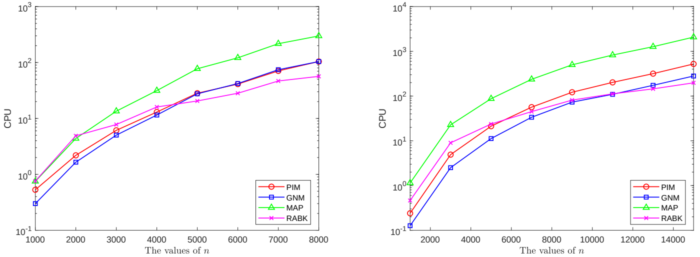


Figure: The averages CPU times for PIM, GNM, MAP, and RABK when using square coefficient matrices. Figures depict the evolution of CPU time vs the increasing dimensions of the coefficient matrices. We have $m = n$, and B is either a random Gaussian matrix (left) or an identity matrix (right).

Comparison of SLA, MAP, and RABK

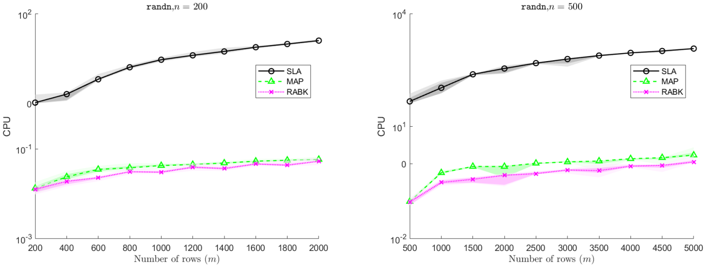


Figure: Comparison of SLA, MAP, and RABK with non-square coefficient matrices. Figures depict the CPU time (in seconds) vs increasing number of rows. The title of each plot indicates the values of n.

Concluding remarks

- We studied the r-set-Douglas-Rachford method enriched with randomization and heavy ball momentum for solving the linear systems.
- We established an adaptive stochastic heavy ball momentum (ASHBM) method for solving consistent linear systems.
- We proposed a simple and versatile randomized iterative algorithmic framework for solving GAVE, applicable to both square and non-square coefficient matrices.
- Nesterov's momentum has gained popularity as a momentum acceleration technique, and recent studies have introduced variants of Nesterov's momentum for accelerating stochastic optimization algorithms. It should be a valuable topic to explore the adaptive Nesterov's momentum.

Thank you for your attention!

https://github.com/xiejx-math