Accelerated Gossip via Stochastic Heavy Ball Method

Nicolas Loizou* & Peter Richtárik*♮†

*University of Edinburgh ♮King Abdullah University of Science and Technology †Moscow Institute of Physics and Technology

1. Average Consensus Problem (ACP)

SETUP: $G = (V, E)$ is a connected network with $|V| = n$ nodes (e.g., sensors) and $|E| = m$ edges (e.g., communication links). Node $i \in V$ stores a private value $c_i \in \mathbb{R}$ (e.g., temperature).

GOAL: Compute the average of the private values (i.e., the quantity $\bar{c} := \frac{1}{n}\sum_i c_i$) in a **distributed** fashion. That is, exchange of information can only occur along the edges.

2. Optimization Formulation of ACP

The optimal solution of the optimization problem

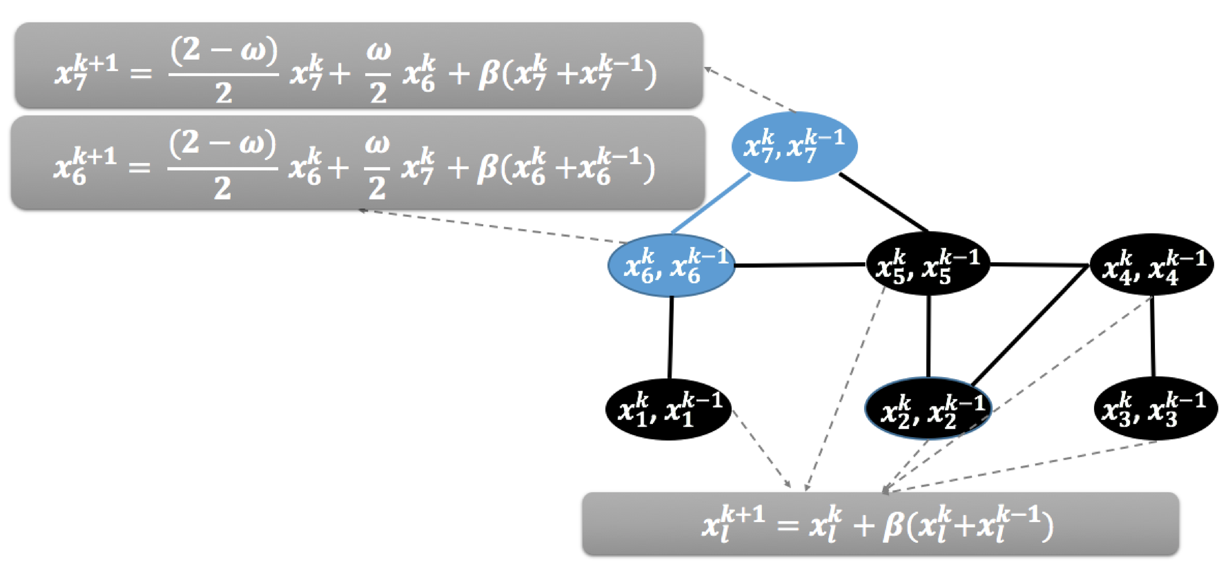
$$\min_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\sum_i (x_i - c_i)^2 \quad \text{subject to} \quad x_i = x_j \quad \text{for all} \quad (i,j) \in E \qquad (1)$$

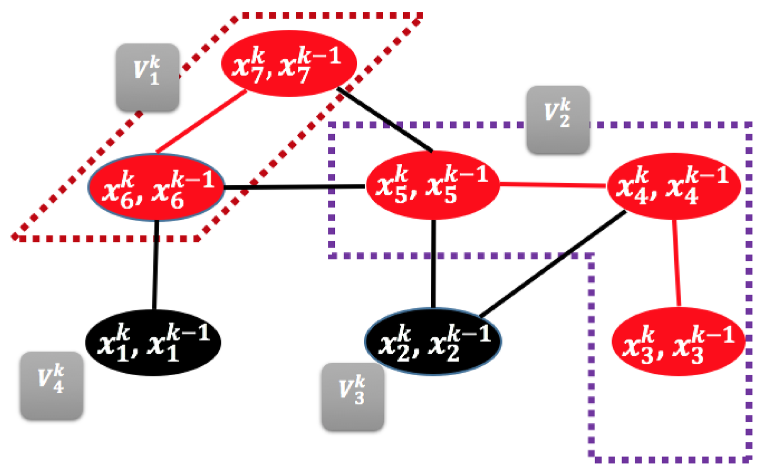
is $x_i^* = \bar{c}$ for all i. The constraints can be written compactly as $\mathbf{A}x = 0$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$, and the rows of the system enforce the constraints $x_i = x_j$ for $(i,j) \in E$.

QUESTIONS: Can we interpret old **RG** algorithms for ACP as instances of specific randomized optimization methods for (1)? Can new **RG** methods be developed this way? Can we developed accelerated **RG** methods?

5. Randomized Kaczmarz Method with Momentum

NEW GOSSIP METHODS: We can now formulate many new variants of **RG**, by applying SHB to (1) with various choices of random matrices $\mathbf{S} \sim \mathcal{D}$.





RK with momentum (mRK):

1. Pick an edge $e = (i,j)$ following the distribution $\mathcal{D}$. In this case $\mathbf{S}_k = e_i$.
2. The values of the nodes are updated as follows:
   - Node i: $x_i^{k+1} = \frac{2-\omega}{2}x_i^k + \frac{\omega}{2}x_j^k + \beta(x_i^k - x_i^{k-1})$
   - Node j: $x_j^{k+1} = \frac{2-\omega}{2}x_j^k + \frac{\omega}{2}x_i^k + \beta(x_j^k - x_j^{k-1})$
   - Any other node ℓ: $x_\ell^{k+1} = x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1})$

Randomized Block Kaczmarz with momentum (mRBK):

1. Form a subgraph G_k of G by selecting a random set of edges $\mathcal{S}_k \subseteq E$. Now $\mathbf{S} = \mathbf{I}_{:C}$ with $C \subseteq [m]$.
2. The values of the nodes are updated as follows: For each connected component $\mathcal{V}_r^k$ of $\mathcal{G}_k$, replace the values of its nodes with:

$$x_i^{k+1} = \omega \left[\frac{\sum_{j \in \mathcal{V}_r^k} x_j^k}{|\mathcal{V}_r^k|}\right] + (1-\omega)x_i^k + \beta(x_i^k - x_i^{k-1})$$

Any other node ℓ: $x_\ell^{k+1} = x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1})$

3. New Viewpoints

Best Approximation Problem:

$$\min_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\sum_i (x_i - c_i)^2 \quad \text{subject to} \quad \mathbf{A}x = b$$

Stochastic Reformulation [7]:

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}_{\mathbf{S} \sim \mathcal{D}}[f_{\mathbf{S}}(x)], \qquad (2)$$

where, $f_{\mathbf{S}}(x) := \frac{1}{2}\|\mathbf{A}x - b\|^2_{\mathbf{H}}$ and $\mathbf{H} := \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top$

4. Stochastic Heavy Ball [5]

Algorithm 1 Stochastic Heavy Ball (SHB)

1. **Parameters:** Distribution $\mathcal{D}$ from which to sample matrices; stepsize/relaxation parameter $\omega > \mathbb{R}$; momentum parameter $\beta \geq 0$.
2. **Initialize:** $x^0, x^1 \in \mathbb{R}^n$
3. **for** $k = 1, 2, \dots$ **do**
4. Draw a fresh $\mathbf{S}_k \sim \mathcal{D}$
5. Set $x^{k+1} = x^k - \omega \nabla f_{\mathbf{S}_k}(x^k) + \beta(x^k - x^{k-1})$
6. **end for**

6. Theoretical Results and Numerical Experiments

L2 **Convergence:**

Theorem 1 [5] Let $\lambda_{\min}^+$ (resp. $\lambda_{\max}$) be the smallest nonzero (resp. largest) eigenvalue of $\mathbf{W} := \mathbf{A}^\top \mathbb{E}[\mathbf{H}]\mathbf{A}$. Assume $0 < \omega < 2$ and $\beta \geq 0$ and that the expressions $a_1 := 1 + 3\beta + 2\beta^2 - (\omega(2-\omega) + \omega\beta)\lambda_{\min}^+$ and $a_2 := \beta + 2\beta^2 + \omega\beta\lambda_{\max}$ satisfy $a_1 + a_2 < 1$. Then

$$\boxed{\mathbb{E}[\|x^k - x^*\|^2] \leq q^k(1+\delta)\|x^0 - x^*\|^2}$$

where $q = \frac{a_1 + \sqrt{a_1^2 + 4a_2}}{2}$ and $\delta = q - a_1$. Moreover, $a_1 + a_2 \leq q < 1$.

L1 **Convergence:**

Theorem 2 [5] Let $0 < \omega \leq 1/\lambda_{\max}$ and $(1 - \sqrt{\omega\lambda_{\min}^+})^2 < \beta < 1$. Then $\exists C > 0$ such that for all $k \geq 0$ we have

$$\boxed{\|\mathbb{E}[x^k - x^*]\|^2 \leq \beta^k C}$$

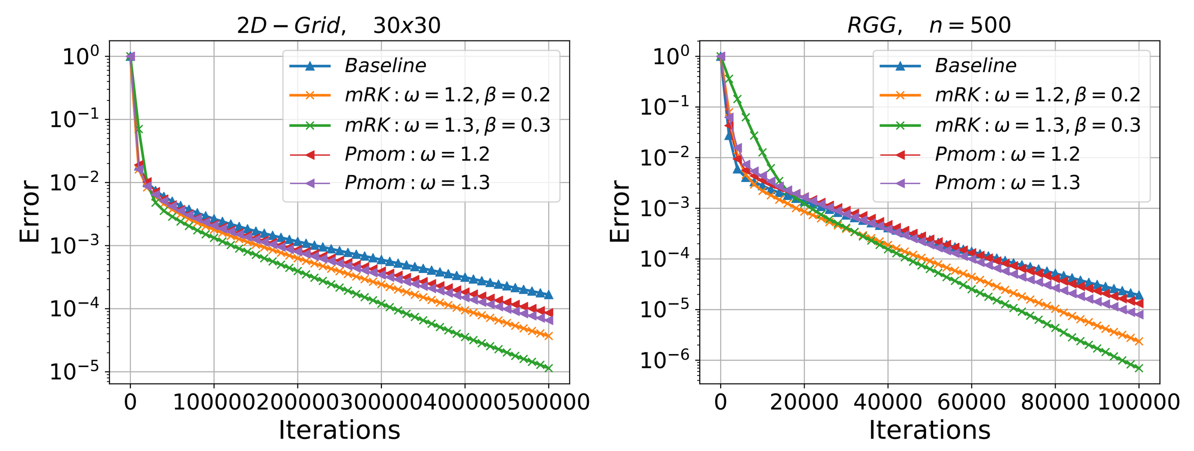


Figure 1: mRK vs. simple pairwise gossip (Baseline) vs. pairwise momentum method (Pmom) [3]

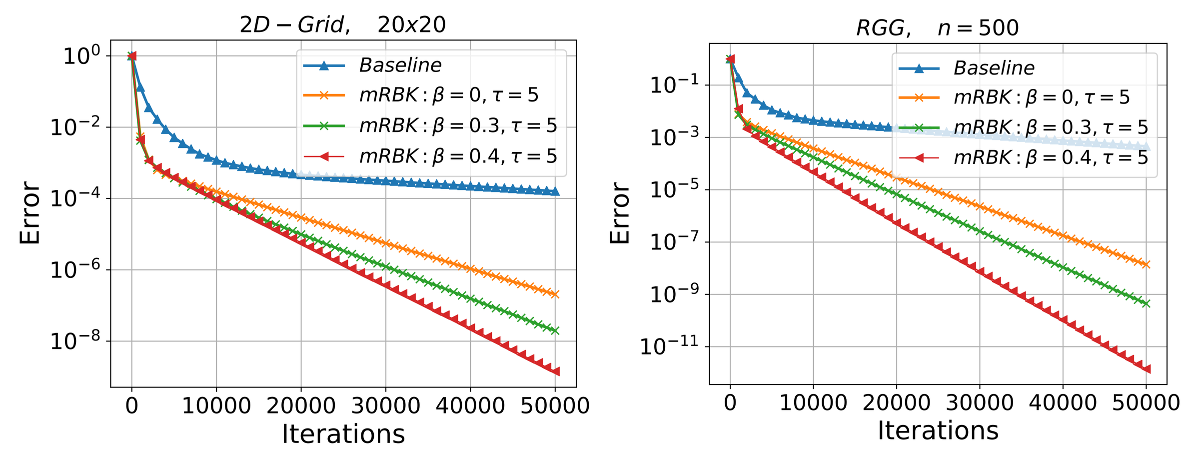


Figure 2: mRBK vs RBK ($\beta = 0$) [4] (stepsize: $\omega = 1$; block size: $\tau = 5$.

7. References

[1] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Trans. Inf. Theory*, 2006.

[2] Robert Mansel Gower and Peter Richtárik. Stochastic dual ascent for solving linear systems. *arXiv:1512.06890*, 2015.

[3] J. Liu, B.D.O Anderson, M. Cao, and A.S. Morse. Analysis of accelerated gossip algorithms. *Automatica*, 2013.

[4] N. Loizou and P. Richtárik. A new perspective on randomized gossip algorithms. In *GlobalSIP*, 2016.

[5] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv:1712.09677*, 2017.

[6] N. Loizou and P. Richtárik. Accelerated gossip via stochastic heavy ball method. In *Allerton*, 2018.

[7] P. Richtárik and M. Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *arXiv:1706.01108*, 2017.