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 \( \bar{c} := \frac{1}{n} \sum c_i \)) 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

\[
\min_{x \in \mathbb{R}^n} \; \frac{1}{2} \sum_{i,j} (x_i - x_j)^2 \quad \text{subject to} \quad x_i = x_j \quad \text{for all} \quad (i,j) \in E
\]

is \( x_i = \bar{c} \) for all \( i \). The constraints can be written compactly as \( Ax = 0 \), where \( A \in \mathbb{R}^{m \times n} \), 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 methods be developed this way? Can we develop 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 \( S \sim \mathcal{D} \).

**RK with momentum (mRK):**

1. Pick an edge \( e = (i,j) \) following the distribution \( \mathcal{D} \). In this case \( S_e = c_i \).
2. The values of the nodes are updated as follows:
   - Node \( i \): \( x_i^{k+1} = \frac{1}{2} x_i^k + \frac{1}{2} \sum_j y_{ij}^{k} + \beta(x_j^k - x_i^k) \)
   - Node \( j \): \( x_j^{k+1} = \frac{1}{2} x_j^k + \frac{1}{2} \sum_i y_{ij}^{k} + \beta(x_j^k - x_i^k) \)
   - Any other node \( \ell \): \( x_{\ell}^{k+1} = x_{\ell}^k + \beta(x_{\ell}^k - x_{\ell}^{k-1}) \)

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 \( W := A^T E [H] A \). Assume \( 0 < \omega < 2 \) and \( \beta \geq 0 \) and that the expressions \( a_1 := 1 + 3\beta + 2\beta^2 - (\omega(2-\omega) + \omega)\lambda_{\min} \) and \( a_2 := \beta + 2\beta^2 + \omega\lambda_{\max} \) satisfy \( a_1 + a_2 < 1 \). Then

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

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

**L1 Convergence:**

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

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

7. References