1. The Problem:

**Stochastic Optimization Problem:**

\[
\min_{x \in \mathbb{R}^d} f(x) := E_{S \sim D}[f_S(x)]
\]

- \( f_S(x) := \frac{1}{2} \|Ax - b\|^2 + \frac{1}{2}(Ax - b)^T H(Ax - b) \) and \( H := S(S^T AA^T S)^T \geq 0 \).
- \( S \) is a random matrix with \( n \) rows (and arbitrary number of columns, e.g., 1).
- \( D \) is a distribution over such matrices.

**Best Approximation Problem:**

\[
\min_{x \in \mathbb{R}^d} P(x) := \frac{1}{2} \|x - x_0\|^2 + \frac{1}{2}(x - x_0)^T(x - x_0)
\]

subject to \( Ax = b \)

**Exactness ([3]):** \( \arg\min_{x \in \mathbb{R}^d} f(x) = \{ x : Ax = b \} \)

2. Stochastic Heavy Ball Method (SHB)

\[
x_{k+1} = x_k - \omega \nabla f_S(x_k) + \beta (x_k - x_{k-1})
\]

- \( S_k \sim D \) in each iteration (i.i.d)
- We do not have (or do not wish to exercise, as it may be prohibitively expensive) explicit access to \( f \). We only have access to stochastic function \( f_S \) and its gradient \( \nabla f_S \).

3. Acceleration mechanism

Let \( S = e_i \) (unit coordinate vector in \( \mathbb{R}^n \)) with probability \( p_i > 0 \). In this setup, SHB simplifies to:

\[
x_{k+1} = x_k - \omega \frac{A^T}{\|A^T\|^2} A x_k + \beta (x_k - x_{k-1})
\]

4. Eigenvalues

\( \lambda_{\max} \) (resp. \( \lambda_{\min}^+ \)) is the largest (resp. smallest nonzero) eigenvalue of \( \nabla^2 f(x) \).

It turns out that \( 0 < \lambda_{\min}^+ \leq \lambda_{\max} \leq 1 \).

5. Convergence Analysis

**L2 Convergence / Function Values**

**Theorem:** Choose \( x_0 = x_1 \in \mathbb{R}^d \). Assume exactness. Let \( \{x_k\}_{k=0}^\infty \) be the sequence of random iterates produced by SHB. Assume \( 0 < \omega \leq 2 \) and \( \beta \geq 0 \) and that the expressions \( a_1 := 1 + 3\beta + 2\beta^2 - (2\omega - \omega + \beta)\lambda_{\max} \) and \( a_2 := \beta + 2\beta^2 + \omega \lambda_{\max} \) satisfy \( a_1 + a_2 < 1 \). Let \( x_k \) be the solution of (2). Then

\[
\mathbb{E}[\|x_k - x_k\|^2] \leq q^k (1 + \delta) \|x_0 - x_k\|^2
\]

**L1 convergence: accelerated linear rate**

**Theorem:** Assume exactness. Let \( \{x_k\}_{k=0}^\infty \) be the sequence of random iterates produced by SHB, started with \( x_0, x_1 \in \mathbb{R}^d \) satisfying the relation \( x_0 - x_1 \in \text{Range}(A^T) \), with stepsize parameter \( 0 < \omega \leq 1/\lambda_{\max} \) and momentum parameter \( (1 - (\omega \lambda_{\max}^+)^2)^2 \beta < 1 \). Then there exists constant \( C > 0 \) such that for all \( k \geq 0 \) we have

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

**Special Cases:**

(i) \( \omega = 1, \beta = (1 - \sqrt{0.99 \lambda_{\max}^+})^2 < 1 \)

\[
\mathbb{E}[\|x_k - x^*\|] \leq 1 - \sqrt{0.99 \lambda_{\max}^+} 2k C
\]

(ii) \( \omega = 1/\lambda_{\max}, \beta = (1 - \sqrt{0.99 \lambda_{\max}})^2 \)

\[
\mathbb{E}[\|x_k - x^*\|] \leq (1 - \sqrt{0.99 \lambda_{\max}}) 2k C
\]

6. Convergence Analysis

**Cesaro average: sublinear rate**

**Theorem:** Choose \( x_0 = x_1 \) and let \( \{x_k\}_{k=0}^\infty \) be the random iterates produced by SHB, where the momentum parameter \( 0 < \beta < 1 \) and relaxation parameter (stepsize) \( \omega \geq 0 \) satisfy \( \omega + 2\beta^2 < 2 \). Let \( x_k \) be any vector satisfying \( f(x_k) = 0 \). If we let \( \bar{x}_k = \frac{1}{k} \sum_{i=1}^{k} x_i \), then

\[
\mathbb{E}[f(\bar{x}_k)] \leq \frac{(1 - \beta)^2 \|x_0 - x^*\|^2 + 2\omega \beta f(\bar{x}_0)}{2\omega(2 - 2\beta - \omega)k}
\]

8. Numerical Evaluation

![Figure 2](image-url)

Figure 2: The performance of randomized Kaczmarz and randomized Kaczmarz with momentum for several momentum parameters \( \beta \) on real data from LIBSVM, mushrooms: \( (n, d) = (8124, 112) \), splice: \( (n, d) = (1000, 60) \). The graphs in the first (second) column plot iterations (time) against residual error while in the third (fourth) column plot iterations (time) against function values. The "Error" on the vertical axis represents the relative error \( \|x_k - x^*\|^2/\|x_0\|^2 \), and the function values \( f(x_k) \) refer to function (1).

9. References