Linearly Convergent Stochastic Heavy Ball Method for Minimizing Generalization Error

Nicolas Loizou* & Peter Richtárik*♮
*University of Edinburgh, ♮KAUST

1. The Problem:

Stochastic Optimization Problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\mathbf{S}\sim\mathcal{D}}[f_{\mathbf{S}}(x)] \tag{1}$$

- $f_{\mathbf{S}}(x) := \frac{1}{2}\|\mathbf{A}x-b\|^2_{\mathbf{H}} = \frac{1}{2}(\mathbf{A}x-b)^\top \mathbf{H}(\mathbf{A}x-b)$ and $\mathbf{H} := \mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S})^\dagger\mathbf{S}^\top \succeq 0$.
- $\mathbf{S}$ is a random matrix with n rows (and arbitrary number of columns, e.g., 1).
- $\mathcal{D}$ is a distribution over such matrices.

Best Approximation Problem:

$$\min_{x\in\mathbb{R}^d} \quad P(x) := \frac{1}{2}\|x-x_0\|^2 := \frac{1}{2}(x-x_0)^\top(x-x_0) \qquad \text{subject to} \quad \mathbf{A}x=b \tag{2}$$

Exactness ([3]) : $\boxed{\operatorname{argmin}_{x\in\mathbb{R}^d} f(x) = \{x : \mathbf{A}x=b\}}$

2. Stochastic Heavy Ball Method (SHB)

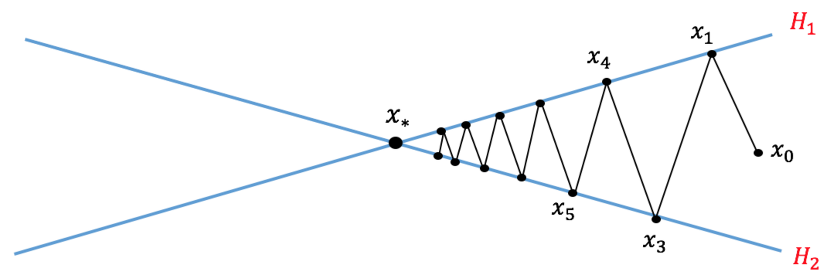
$$\boxed{x_{k+1} = \underbrace{x_k - \omega\nabla f_{\mathbf{S}_k}(x_k)}_{\text{Stochastic Gradient Descent}} \underbrace{+\beta(x_k-x_{k-1})}_{\text{Momentum Term}}}$$

- $\mathbf{S}_k \sim \mathcal{D}$ 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 function f. We only have access to stochastic function $f_{\mathbf{S}_k}$ and its gradient $\nabla f_{\mathbf{S}_k}$.

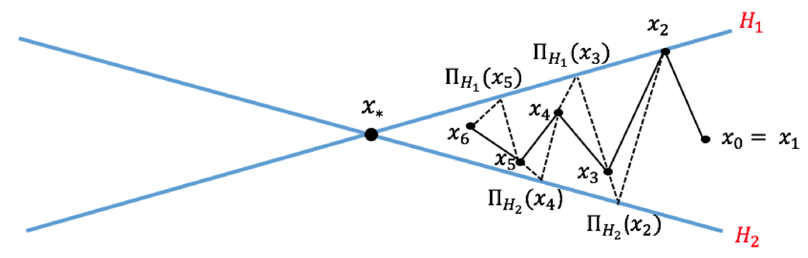
3. Acceleration mechanism

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

$$\boxed{x_{k+1} = x_k - \omega\frac{\mathbf{A}_{i:}x_k-b_i}{\|\mathbf{A}_{i:}\|_2^2}\mathbf{A}_{i:}^\top + \beta(x_k-x_{k-1})}$$



(a) Randomized Kaczmarz Method [4]



(b) Randomized Kaczmarz Method with Momentum [This paper]

Figure 1: Graphical interpretation of the *randomized Kaczmarz method* and the *randomized Kaczmarz method with momentum* on a simple example with only two hyperplanes $H_i=\{x : \mathbf{A}_{i:}x-b_i\}$ where $i=1,2$ and a unique solution x_*. That is, $n=2$ and $d=2$.

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 $0<\omega<2$ 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 $a_1+a_2<1$. Let x_* be the solution of (2). Then

$$\boxed{\mathbb{E}[\|x_k-x_*\|^2] \leq q^k(1+\delta)\|x_0-x_*\|^2} \tag{3}$$

$$\boxed{\mathbb{E}[f(x_k)] \leq q^k\frac{\lambda_{\max}}{2}(1+\delta)\|x_0-x_*\|^2} \tag{4}$$

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

4. Eigenvalues

$\lambda_{\max}$ (resp. $\lambda_{\min}^+$) is the largest (resp. smallest nonzero) eigenvalue of $\nabla^2 f(x)$.
It turns out that $0<\lambda_{\min}^+\leq\lambda_{\max}\leq 1$.

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 $0\leq\beta<1$ and relaxation parameter (stepsize) $\omega\geq 0$ satisfy $\omega+2\beta<2$. Let x_* be any vector satisfying $f(x_*)=0$. If we let $\hat{x}_k=\frac{1}{k}\sum_{t=1}^k x_t$, then

$$\boxed{\mathbb{E}[f(\hat{x}_k)] \leq \frac{(1-\beta)^2\|x_0-x_*\|^2+2\omega\beta f(x_0)}{2\omega(2-2\beta-\omega)k}} \tag{5}$$

7. Convergence Analysis

***L1* convergence: accelerated linear rate**

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

$$\boxed{\|\mathbb{E}[x_k-x_*]\|^2 \leq \beta^k C} \tag{6}$$

Special Cases:

(i) $\omega=1,\ \beta=\left(1-\sqrt{0.99\lambda_{\min}^+}\right)^2$:

$$\|\mathbb{E}[x_k-x_*]\|^2_{\mathbf{B}} \leq \left(1-\sqrt{0.99\lambda_{\min}^+}\right)^{2k} C$$

(ii) $\omega=1/\lambda_{\max},\ \beta=\left(1-\sqrt{0.99\frac{\lambda_{\min}^+}{\lambda_{\max}}}\right)^2$:

$$\|\mathbb{E}[x_k-x_*]\|^2_{\mathbf{B}} \leq \left(1-\sqrt{0.99\frac{\lambda_{\min}^+}{\lambda_{\max}}}\right)^{2k} C$$

8. Numerical Evaluation

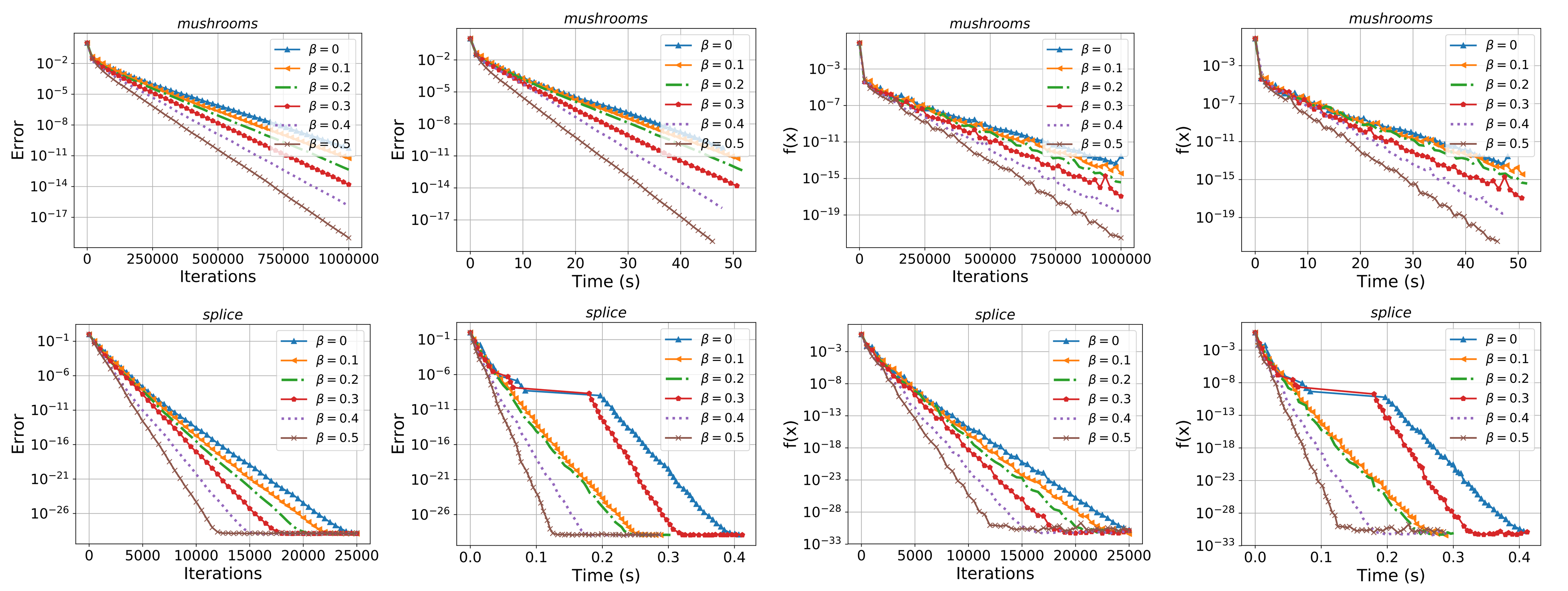


Figure 2: The performance of randomized Kaczmarz and randomized Kaczmarz with momentum for several momentum parameters β on real data from LIBSVM. mushrooms: $(n,d)=(8124,112)$, splice: $(n,d)=(1000,60)$. The graphs in the first (second) column plot iterations (time) against residual error while in the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

9. References

[1] N. Loizou and P. Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *NIPS-OPT workshop*, 2017.

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

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

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