Convergence of a Stochastic Gradient Method with Momentum for Non-Smooth Non-Convex Optimization

Vien V. Mai and Mikael Johansson
KTH - Royal Institute of Technology

Stochastic optimization

Stochastic optimization problem:

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x; S)] = \int_{\mathcal{S}} f(x; s) dP(s)$$

Stochastic gradient descent (SGD):

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k, S_k)$$

SGD with momentum:

$$x_{k+1} = x_k - \alpha_k z_k, \qquad z_{k+1} = \beta_k g_{k+1} + (1 - \beta_k) z_k$$

Includes Polyak's Heavy ball, Nesterov's fast gradient, and more

- widespread empirical success
- theory less clear than deterministic counterpart

Stochastic optimization: sample complexity

For SGD, sample complexity is known under various assumptions

- **convexity** [Nemirovski et al., 2009]
- **smoothness** [Ghadimi-Lan, 2013]
- **weak convexity** [Davis-Drusvyatskiy, 2019]

Much less is known for momentum-based methods

- constrained
- non-smooth non-convex

Our contributions

Novel Lyapunov analysis for (projected) stochastic heavy ball (SHB):

- sample complexity of SHB for stochastic weakly convex minimization
- analyze smooth non-convex case under less restrictive assumptions

Outline

- Background and motivation
- **SHB for non-smooth non-convex optimization**
- Sharper results for smooth non-convex optimization
- Numerical examples
- Summary and conclusions

Problem formulation

Problem:

$$\underset{x\in\mathcal{X}}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s)$$

$\mathcal{X}$ is closed and convex; f is ρ-**weakly convex**, meaning that

$$x \mapsto f(x) + \rho\left\|x\right\|_2^2 \text{ is convex.}$$

Easy to recognize, e.g., convex compositions

$$f(x) = h(c(x))$$

h convex and L_h-Lipschitz; c smooth with L_c-Lipschitz Jacobian ($\rho = L_h L_c$)

Algorithm

Consider

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s)$$

Algorithm:

$$x_{k+1} = \underset{x \in \mathcal{X}}{\text{argmin}} \left\{ \langle z_k, x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

$$z_{k+1} = \beta g_{k+1} + (1-\beta)\frac{x_k - x_{k+1}}{\alpha}$$

Recovers SHB when $\mathcal{X} = \mathbb{R}^n$; setting $\beta = 1$ gives (projected) SGD

Goal: establish sample complexity

Roadmap and challenges

Most complexity results for subgradient-based methods rely on forming:

$$\mathbb{E}[V_{k+1}] \le \mathbb{E}[V_k] - \alpha\,\mathbb{E}[e_k] + \alpha^2 C^2$$

Immediately yields $O(1/\epsilon^2)$ complexity for $\mathbb{E}[e_k]$

Stationarity measure:

- f convex $\Longrightarrow e_k = f(x_k) - f(x^\star)$; f smooth $\Longrightarrow e_k = \|\nabla f(x_k)\|_2^2$
- f weakly convex $\Longrightarrow e_k = \|\nabla F_\lambda(x_k)\|_2^2$

Lyapunov analysis (for SGD):

- f convex $\Longrightarrow V_k = \|x_k - x^\star\|_2^2$ [Shor, 1964]
- f smooth $\Longrightarrow V_k = f(x_k)$ [Ghadimi-Lan, 2013]
- f weakly convex $\Longrightarrow V_k = F_\lambda(x_k)$ [Davis-Drusvyatskiy, 2019]

Convergence to stationarity in weakly convex cases

Moreau envelope

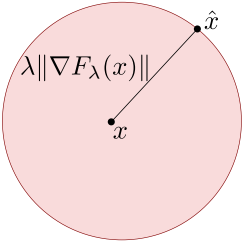
$$F_\lambda(x) = \inf_y \left\{ F(y) + \frac{1}{2\lambda} \|x - y\|_2^2 \right\}$$

Proximal mapping

$$\hat{x} := \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ F(y) + \frac{1}{2\lambda} \|x - y\|_2^2 \right\}$$

Connection to near-stationarity

$$\begin{cases} \lambda^{-1}(x - \hat{x}) = \nabla F_\lambda(x) \\ \mathrm{dist}(0, \partial F(\hat{x})) \le \|\nabla F_\lambda(x)\|_2 \end{cases}$$



Small $\|\nabla F_\lambda(x)\|_2 \implies x$ close to a near-stationary point

Lyapunov analysis for SHB

Recall that we wanted

$$\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] - \alpha\, \mathbb{E}[e_k] + \alpha^2 C^2$$

SGD works with $e_k = \|\nabla F_\lambda(x_k)\|_2^2$ and $V_k = F_\lambda(x_k)$

It seems natural to take $e_k = \|\nabla F_\lambda(\cdot)\|_2^2$

Two questions:

- at which point should we evaluate $\nabla F_\lambda(\cdot)$?
- can we find a corresponding Lyapunov function V_k?

Lyapunov analysis for SHB

Our approach: Take $\nabla F_\lambda(\cdot)$ at the following iterate

$$\bar{x}_k := x_k + \frac{1-\beta}{\beta}(x_k - x_{k-1})$$

Define the corresponding proximal point

$$\hat{x}_k = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ F(y) + \frac{1}{2\lambda} \|y - \bar{x}_k\|_2^2 \right\}$$

This gives

$$\boxed{e_k = \nabla F_\lambda(\bar{x}_k) = \lambda^{-1}(\bar{x}_k - \hat{x}_k)}$$

Lyapunov analysis for SHB

Let $\beta = \nu\alpha$ so that $\beta \in (0, 1]$ and define $\xi = (1 - \beta)/\nu$.

Consider the function:

$$V_k = F_\lambda(\bar{x}_k) + \frac{\nu\xi^2}{4\lambda^2}\|p_k\|_2^2 + \frac{\alpha\xi^2}{2\lambda^2}\|d_k\|_2^2 + \left(\frac{(1-\beta)\xi^2}{2\lambda^2} + \frac{\xi}{\lambda}\right) f(x_{k-1}),$$

where

$$p_k = \frac{1-\beta}{\beta}\left(x_k - x_{k-1}\right) \text{ and } d_k = \left(x_{k-1} - x_k\right)/\alpha.$$

Theorem: For any $k \in \mathbb{N}$, it holds that

$$\mathbb{E}\left[V_{k+1}\right] \leq \mathbb{E}[V_k] - \frac{\alpha}{2}\mathbb{E}[\|\nabla F_\lambda(\bar{x}_k)\|_2^2] + \frac{\alpha^2 C L^2}{2\lambda}.$$

Main result: sample complexity

Taking $\alpha = \alpha_0/\sqrt{K}$ and $\beta = O(1/\sqrt{K}) \in (0, 1]$ yields

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(\bar{x}_{k^*})\right\|_2^2\right] \leq O\left(\frac{\rho\Delta + L^2}{\sqrt{K+1}}\right)$$

$\Delta = f(x_0) - \inf_{x\in\mathcal{X}} f(x)$

Note:

- same worst-case complexity as SGD ($\beta = 1$)
- β can be as small as $O(1/\sqrt{K})$
- (much) more weight to the momentum term than the fresh subgradient

This rate is, in general, not possible to improve [Arjevani et al., 2019].

Outline

- Background and motivation
- SHB for non-smooth non-convex optimization
- **Sharper results for smooth non-convex optimization**
- Numerical examples
- Summary and conclusions

Smooth and non-convex optimization

Problem:

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s)$$

$\mathcal{X}$ is closed and convex; f is ρ-**smooth**:

$$\|\nabla f(x) - \nabla f(x)\|_2 \leq \rho \|x - y\|_2\,, \;\; \forall x, y \in \operatorname{dom} f.$$

Assumption. There exists a real $\sigma > 0$ such that for all $x \in \mathcal{X}$:

$$\mathbb{E}\left[\|f'(x,S) - \nabla f(x)\|_2^2\right] \leq \sigma^2.$$

Note.

- complexity of SHB is not known (even for deterministic case)
- when $\mathcal{X} = \mathbb{R}^n$, $O(1/\epsilon^2)$ obtained under bounded gradients assumption

[Yan et al., 2018]

Improved complexities on smooth non-convex problems

Constrained case:

Suppose that $\|\nabla f(x)\|_2 \leq G$ for all $x \in \mathcal{X}$. If we set $\alpha = \frac{\alpha_0}{\sqrt{K+1}}$, then

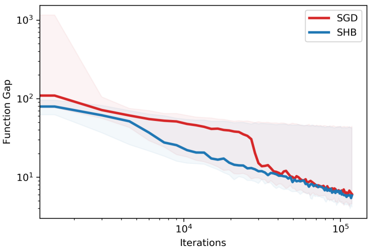
$$\mathbb{E}\left[\|\nabla F_\lambda(\bar{x}_{k^*})\|_2^2\right] \leq O\left(\frac{\rho\Delta + \sigma^2 + G^2}{\sqrt{K+1}}\right).$$

Unconstrained case:

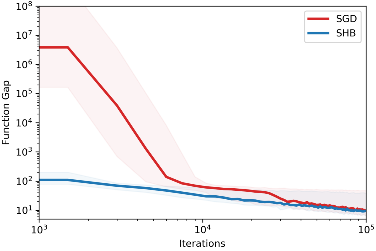
If we set $\alpha = \frac{\alpha_0}{\sqrt{K+1}}$ with $\alpha_0 \in (0, 1/(4\rho)]$, then

$$\mathbb{E}\left[\|\nabla F_\lambda(\bar{x}_{k^*})\|_2^2\right] \leq O\left(\frac{\left(1 + 8\rho^2\alpha_0^2\right)\Delta + (\rho + 16\alpha_0\rho^2)\sigma^2\alpha_0^3}{\alpha_0\sqrt{K+1}}\right).$$

Experiments: convergence behavior on phase retrieval



(a) $\kappa = 1,\ \alpha_0 = 0.1$



(b) $\kappa = 1,\ \alpha_0 = 0.15$

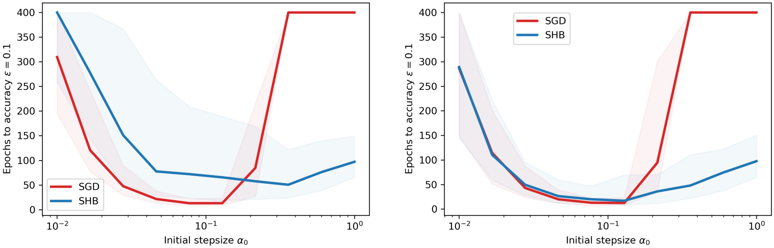
Figure: Function gap vs. #iters for phase retrieval with $p_{\text{fail}} = 0.2,\ \beta = 10/\sqrt{K}$.

Exponential growth before eventual convergence[1] not shown

SGD is competitive if well-tuned, but sensitive to stepsize choice

[1] observed also in [Asi-Duchi, 2019]

Experiments: sensitivity to initial stepsize

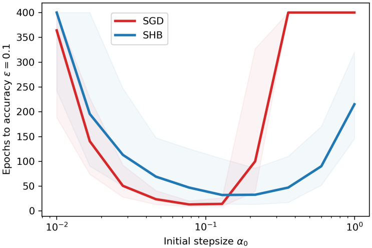


(a) $\beta = 1/\sqrt{K}$

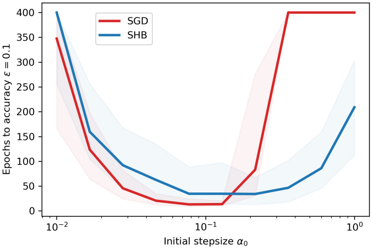
(b) $\beta = 1/\alpha_0/\sqrt{K}$

Figure: #epochs to achieve ϵ-accuracy vs. initial stepsize α_0 with $\kappa = 10$.

Experiments: popular momentum parameter



(a) $1 - \beta = 0.9$



(b) $1 - \beta = 0.99$

Figure: #epochs to achieve ϵ-accuracy vs. initial stepsize α_0 with $\kappa = 10$.

Conclusion

SGD with momentum

- simple modifications to SGD
- good performance and less sensitive to algorithm parameters

Novel Lyapunov analysis

- sample complexity of SHB for weakly convex and constrained optim.
- improved rates on smooth and non-convex problems