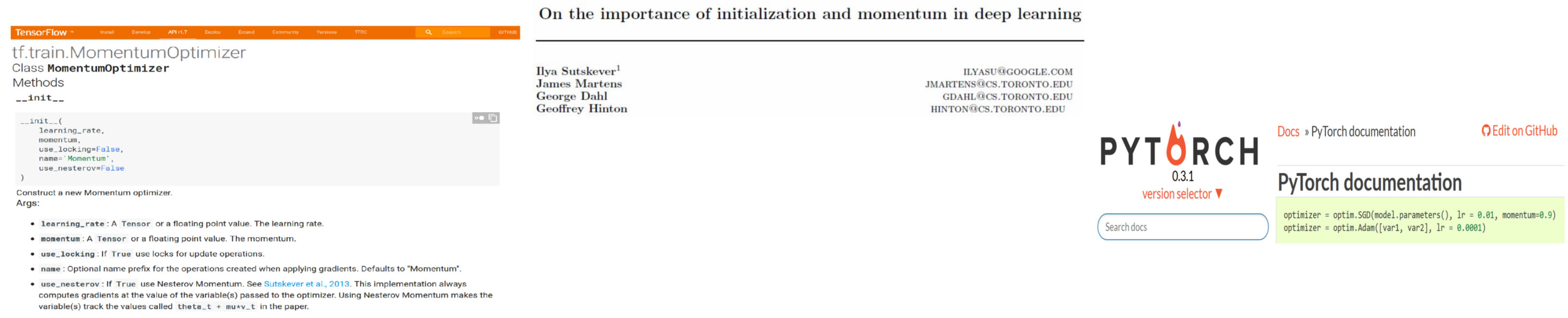
On The Insufficiency Of Existing Momentum Schemes For Stochastic Optimization

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain and Sham M. Kakade

University of Washington, Seattle; Microsoft Research, India

Momentum really works



- Deeply influential: SGD means SGD + Momentum.
- Rigorous understanding lacking.
- This work: initiates understanding of Heavy Ball (HB) Momentum [Polyak, 1964] and Nesterov's Acceleration (NAG) [Nesterov, 1983] with stochastic gradients.

Problem setup and folklore results

- \`\`**n**'' examples: $(x_1, y_1), \cdots (x_n, y_n) \sim D(R^d \times R)$
- **Ultimate Goal**: $w^* = \arg\min L(w) = E[(y - \langle w, x\rangle)^2]$.

$$H = 2 \cdot E[xx^\top]; \quad \kappa_{GD} = \frac{\lambda_{max}[H]}{\lambda_{min}[H]}; \quad \kappa = \frac{\max ||x||^2}{\lambda_{min}(H)}.$$

- GD: $\Theta(\kappa_{GD} \log 1/\epsilon)$ iterations.
- NAG/HB: $\Theta(\sqrt{\kappa_{GD}} \log 1/\epsilon)$ iterations.
- SGD [Jain et al. 2016]: $\Theta(\kappa \log 1/\epsilon)$ iterations.
  - ➢ Assumes realizable model: $y = \langle w^*, x\rangle$.
  - ➢ Applicable to general *agnostic* case. Refer to paper.

Stochastic HB doesn't improve on SGD

- Rigorous proof with an example

HB + Stochastic Gradients requires $\Omega(\kappa \log 1/\epsilon)$ iterations.

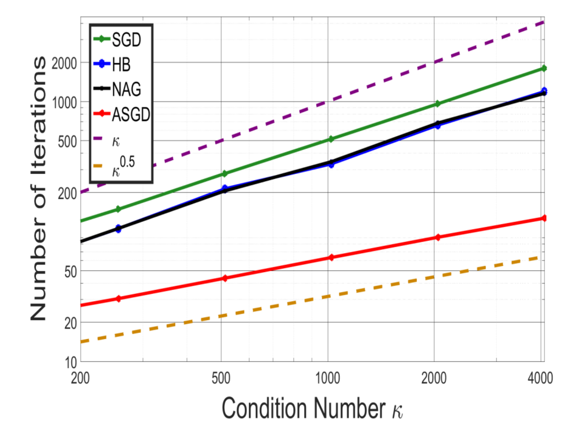
- Empirically, appears true for Gaussian inputs.
- Empirically, lower bound holds for NAG.

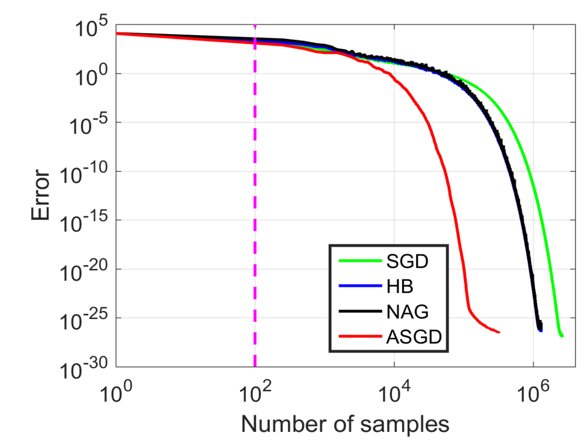
HB/NAG: No improvement over SGD on generic instances.

- This result is **not** a worst case characterization!

Provably improving on SGD

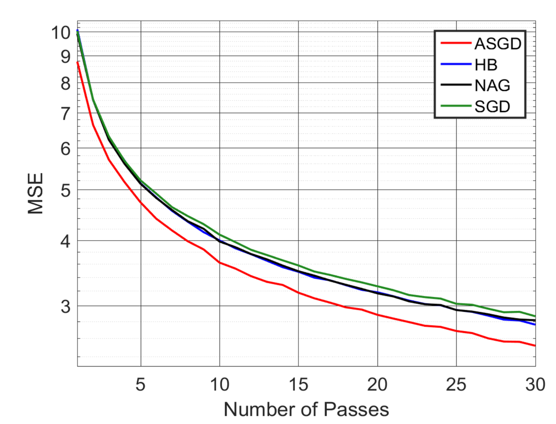
- Gaussian inputs: HB/NAG no speedups on SGD.
- Accelerated SGD [Jain, Kakade, Kidambi, Netrapalli, Sidford 2017]: $\tilde{O}(\sqrt{\kappa d} \log 1/\epsilon)$.

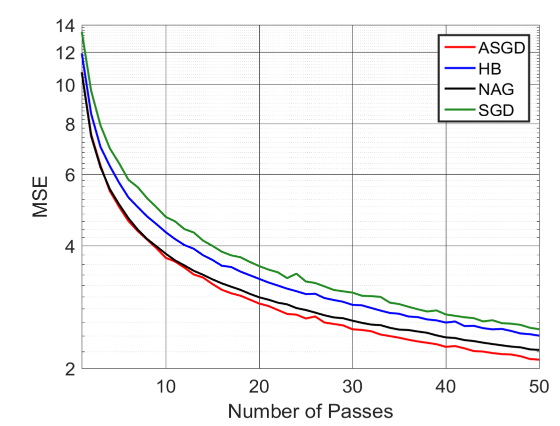




Empirical validation

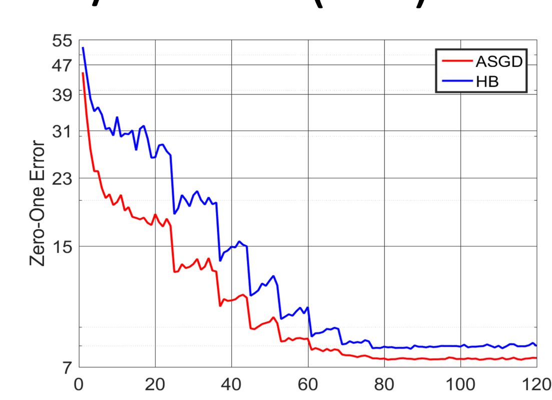
MNIST Autoencoder: batch size 1(left), 8(right).

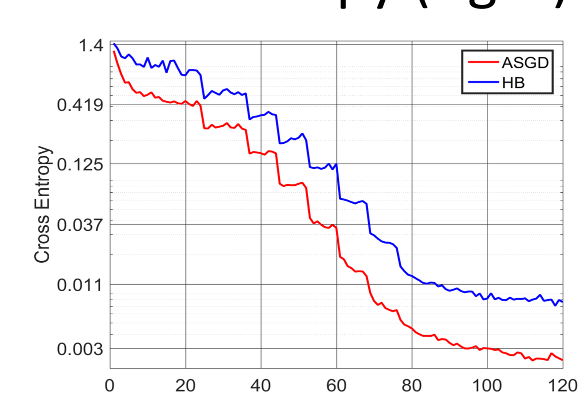


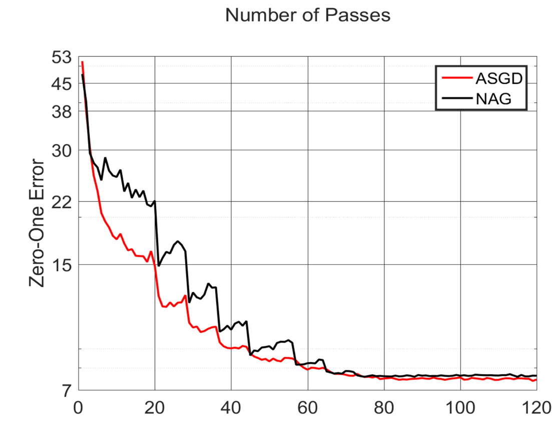


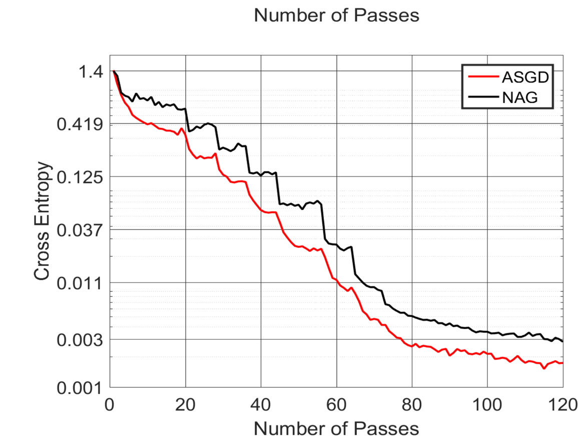
Resnet-44 [He et al. 16] on CIFAR-10 with batch size 128:

- Exhaustive grid search comparing HB/NAG with ASGD.
- Test 0/1 Error (left) and Train Cross-Entropy (right).



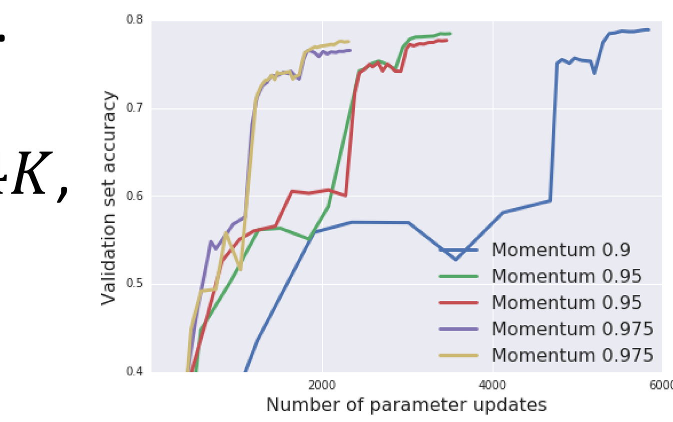






Why momentum still works in practice?

- Sole reason: **Mini-batching**!
- Stochastic gradient → exact gradient.
- Smith et al., ICLR 2018: \`\`increased batch size allows using larger momentum''. See right.
- Batch size used in their work:
  - ➢ Blue $\approx 8K$, red/green $\approx 14K$, purple/yellow $\approx 19K$.



Broader perspective(s)

- Classical optimization: immense practical impact.
- Sharp theory often lacking.
- Rethink large-scale learning [Bottou & Bousquet'08] using stochastic approximation.
- Goal: understand and improve SGD on per-problem basis.
  1. Jain, Kakade, Kidambi, Netrapalli, Sidford 2016: understands SGD's parallelization properties.
  2. Jain, Kakade, Kidambi, Netrapalli, Sidford 2017: first method (ASGD) provably faster than SGD.
- **Plenty** of impactful questions open.

Concluding remarks

- Be(a)ware of employing deterministic optimization methods with stochastic gradients.
- Exciting speedups observed due to mini-batching.
- Significant gains using dedicated stochastic methods.
- Accelerated SGD is the only one known algorithm.
- Many such algorithms/insights still required.

UNIVERSITY of WASHINGTON

ADSI Algorithmic Foundations for Data Science Institute

Microsoft® Research