Simple and optimal high probability bounds for strongly-convex stochastic gradient descent

Nicholas J. A. Harvey Christopher Liaw Sikander Randhawa
University of British Columbia

Motivation:

Standard SGD Results: *non-smooth and strongly convex functions*

SGD, in a nutshell: $x_{t+1} \leftarrow x_t - \eta_t \widehat{g_t}.$ $[\ \mathbb{E}\widehat{g_t} = \nabla f(x_t)\ ]$

Final iterate expected error: $\mathbb{E}[f(x_T) - f(x^*)] = \Theta\left(\frac{\log T}{T}\right)$

[Step size $\eta_t = {}^1/_t.$]

Uniform average expected error: $\mathbb{E}\left[f\left(\frac{1}{T}\sum_1^T x_t\right) - f(x^*)\right] = \Theta\left(\frac{\log T}{T}\right)$

[Step size $\eta_t = {}^1/_t.$]

Lower bound on expected error for any first order, stochastic alg: $\Omega\left(\frac{1}{T}\right)$

Closing the gap: *Getting the optimal expected $\mathcal{O}\left(\frac{1}{T}\right)$ rate.*

Researchers have designed algorithms to get the optimal rate.

Some algorithms are simpler than others.

Arguably, the ***simplest and easiest to implement*** is the following:

- Run SGD *(step size $\eta_t = {}^{2}/_{t+1}$)* until output time.
- Return a ***non-uniform average***:

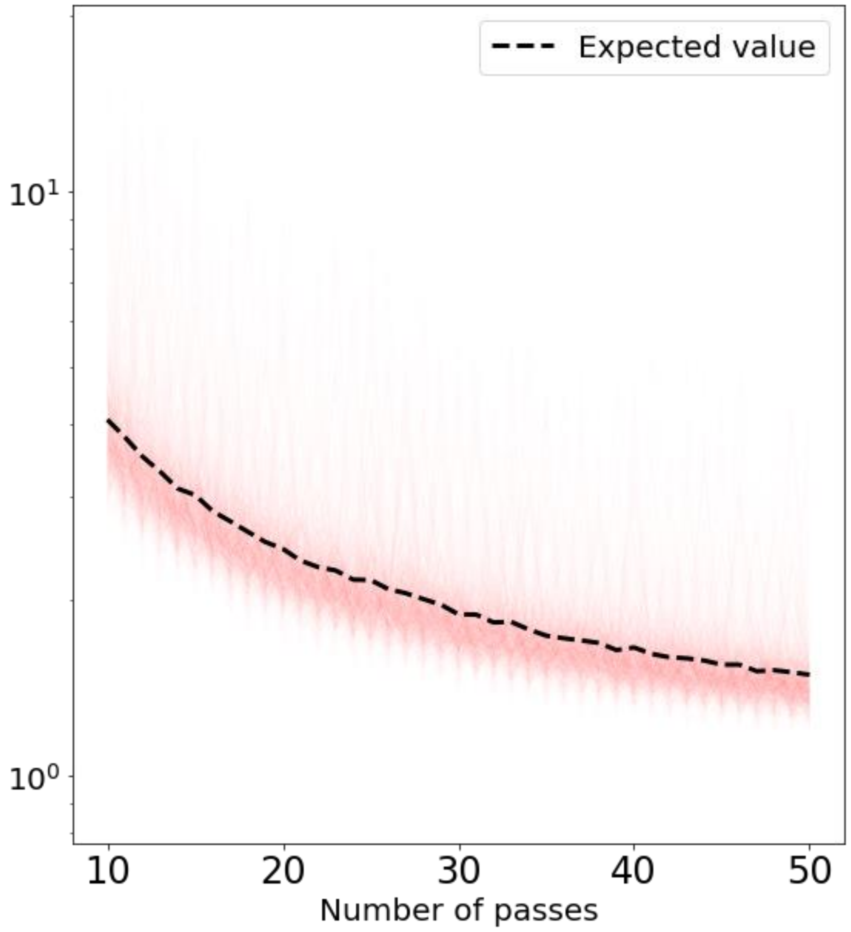
$$\frac{1}{\binom{T}{2}} \sum_{t=1}^{T} t\, x_t.$$

[Lacoste-Julien, Schmidt, Bach (2012)]

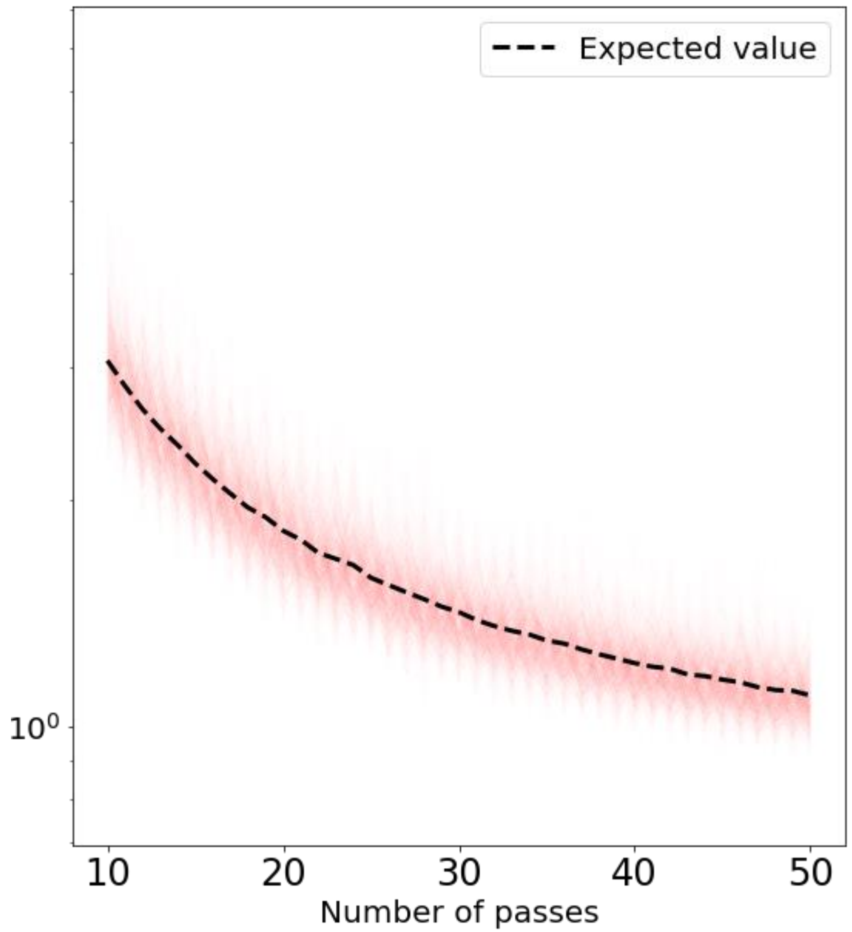
Another Issue: *High variance*

Minimize $f(w) = \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{m} \max\{0, 1 - y_i w^T x_i\}$

Objective value of each iterate of SGD over time (1000 trials)



Protein dataset from KDD world cup 2004



Quantum dataset from KDD world cup 2004

Another Issue: *High variance*

Maybe the non-uniform averaging strategy also suffers from high variance?

Main Result: *High probability bound on error of non uniform average*

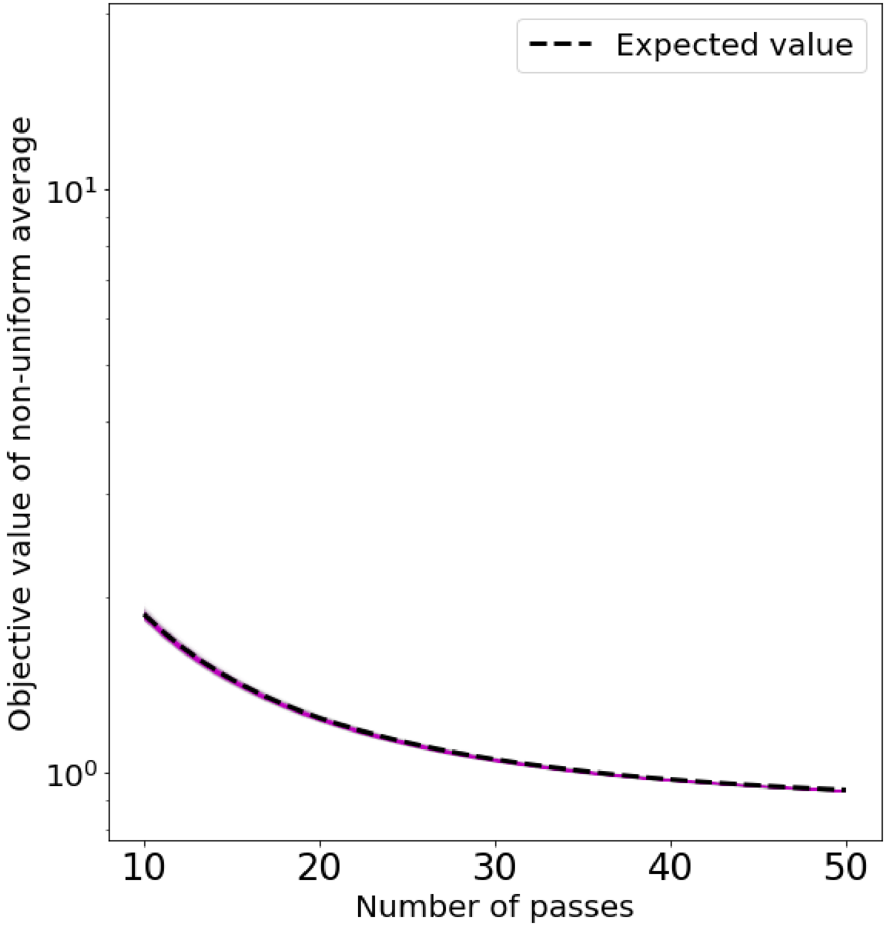
Theorem: Suppose f is strongly-convex and $||\widehat{g_t}||$ is bounded for all t. Run SGD *(with step size $\eta_t = {}^2/_{t+1}$)*. Then, for every $\delta \in (0,1)$,

$$f\left(\frac{1}{\binom{T}{2}}\sum_{t=1}^{T} t x_t\right) - f(x^*) \leq \mathcal{O}\left(\frac{\log\left({}^1/_\delta\right)}{T}\right),$$

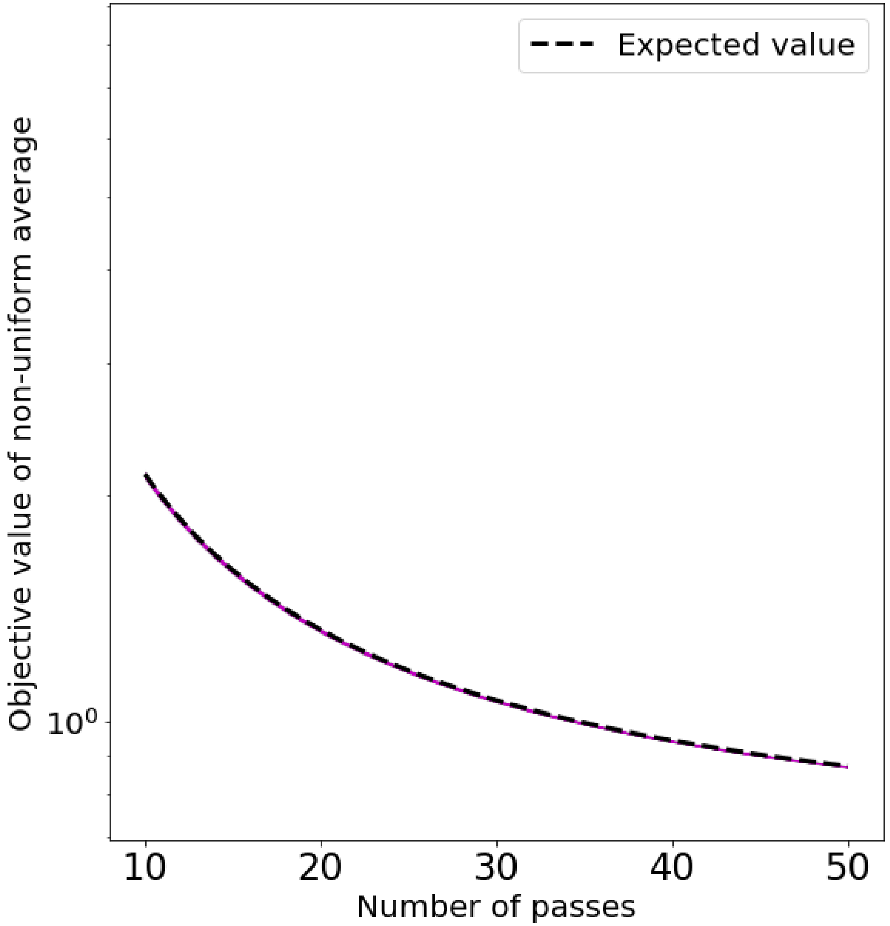
with probability at least $1 - \delta$.

Empirical Performance: *non-uniform average*

Objective value of non-uniform average over time (1000 trials)



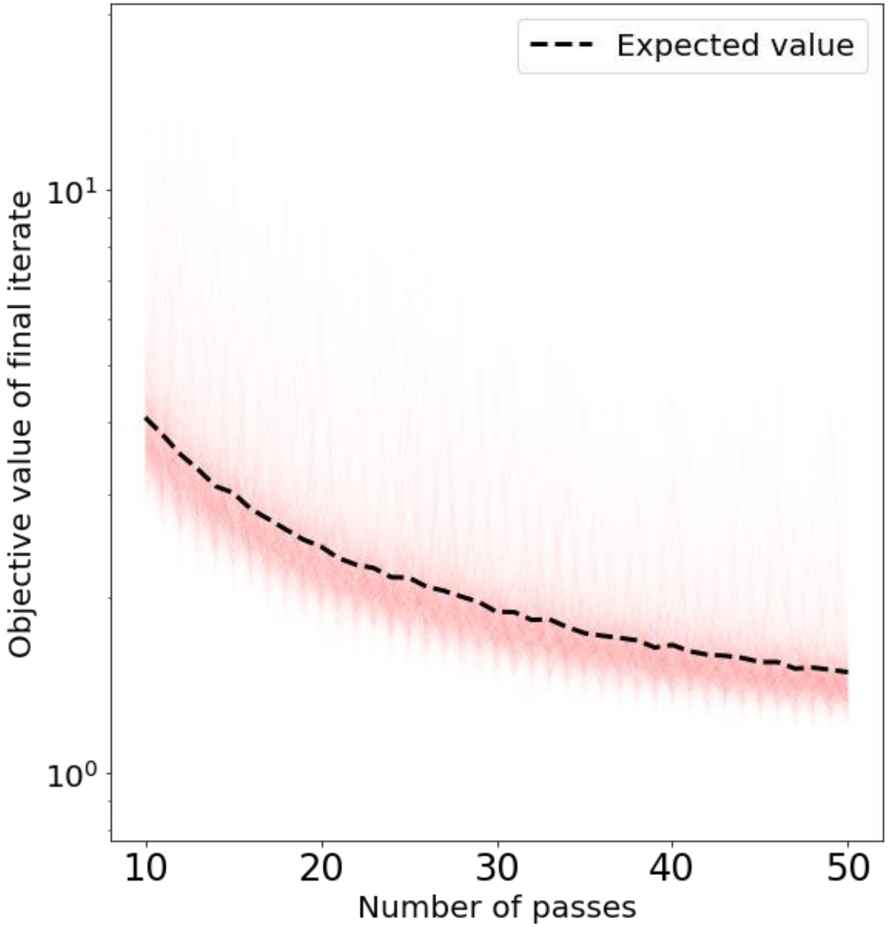
Protein dataset from KDD world cup 2004



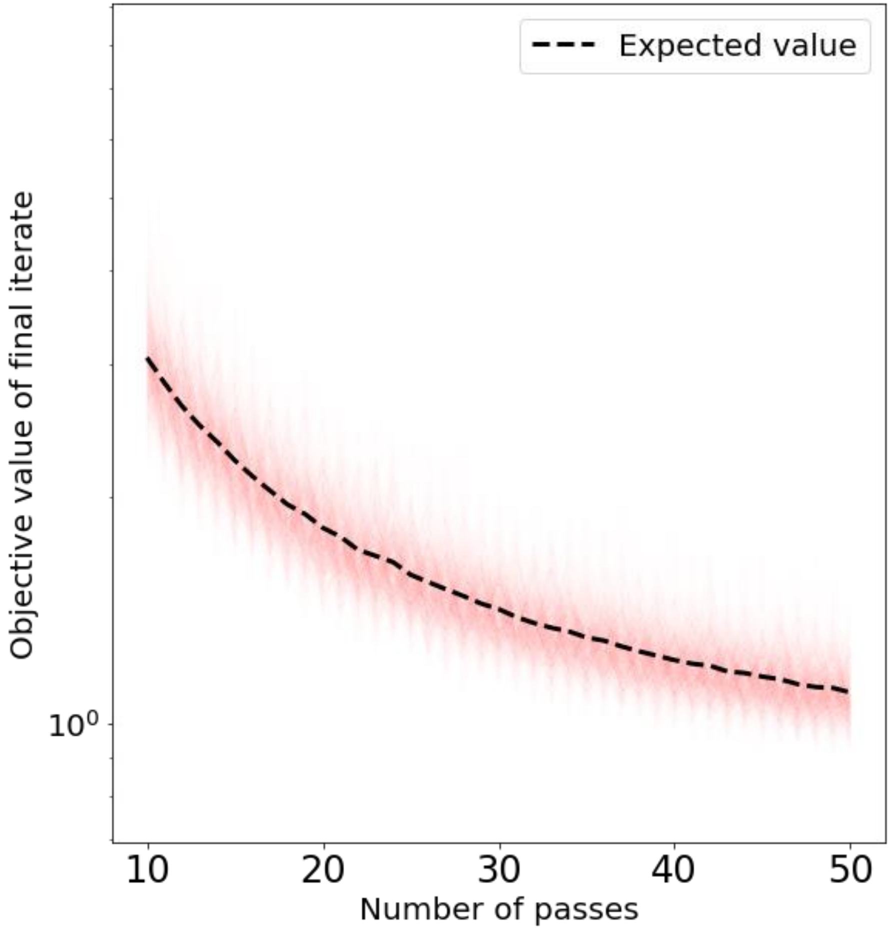
Quantum dataset from KDD world cup 2004

Empirical Performance: *individual iterates*

Objective value of each iterate of SGD over time (1000 trials)



Protein dataset from KDD world cup 2004



Quantum dataset from KDD world cup 2004

Thank you!