Erratum: Second-Order Stochastic Optimization for Machine Learning in Linear Time

Naman Agarwal namana@cs.princeton.edu
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

Brian Bullins bbullins@cs.princeton.edu
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

Elad Hazan ehazan@cs.princeton.edu
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

Editor: Tong Zhang

An error is present in Algorithm 4 and the proof of Theorem 15 in Section 5 of the original manuscript, as a result of an incorrect handling of the quadratic model and its conditioning properties. Thus, we provide in this erratum a correction to this error. First, we amend the bullet points in Section 5.1 to now say:

- Given A we will compute a low complexity constant spectral approximation B of A. Specifically, $B = \sum_{i=1}^{O(d\log(d))} \mathbf{u}_i\mathbf{u}_i^T$ and $\frac{1}{2}B \preceq A \preceq 2B$. This is achieved by techniques developed in matrix sampling/sketching literature, especially those of Cohen et al. (2015). The procedure requires solving a constant number of $O(d\log(d))$ sized linear systems, which we do via Accelerated SVRG.
- We then observe that the quadratic function in A is $\frac{1}{2}$-strongly convex and 2-smooth w.r.t. $\|\cdot\|_B$ (and thus has constant condition number), at which point we may follow the standard descent analysis, accounting for the approximation error incurred when approximately solving a system in B.

Next, we present the corrected versions of Algorithm 4 and the proof of Theorem 15.

Proof [Proof of Theorem 15 (Corrected)] We may first observe that $W(\tilde{\mathbf{v}})$ (defined in Algorithm 4) is $\frac{1}{2}$-strongly convex and 2-smooth with respect to the norm given by $\|\tilde{\mathbf{v}}\|_B \triangleq \sqrt{\tilde{\mathbf{v}}^\top B\tilde{\mathbf{v}}}$. In this case, it is well-known that running an iterative method of the form

$$\tilde{\mathbf{v}}_{t+1} = \tilde{\mathbf{v}}_t - \frac{1}{4}B^{-1}\nabla W(\tilde{\mathbf{v}}_t) \tag{1}$$

will converge to an ε-approximate minimizer of $W(\tilde{\mathbf{v}})$ in $O(\log(h_0/\varepsilon))$ iterations, where $h_0 \triangleq W(\tilde{\mathbf{v}}_0) - \min_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}})$. Thus, all that is left is to handle the approximation error incurred by Acc-SVRG.

Algorithm 4 Fast Quadratic Solver (FQS) (Corrected)

Input: $A = \sum_{i=1}^{m}(\mathbf{v}_i\mathbf{v}_i^T + \lambda I)$**,** $\mathbf{b}$**,** $\varepsilon > 0$**,** $K = \tilde{O}(\log(1/\varepsilon))$**,** $\tilde{\mathbf{v}}_0 = 0$
Output : $\tilde{\mathbf{v}}_K$ **s.t.** $\|A^{-1}\mathbf{b} - \tilde{\mathbf{v}}_K\| \leq \varepsilon$
Compute B s.t. $2B \succeq A \succeq \frac{1}{2}B$ using REPEATED HALVING (Algorithm 3)
Define $W(\tilde{\mathbf{v}}) = \frac{1}{2}\tilde{\mathbf{v}}^\top A\tilde{\mathbf{v}} - \mathbf{b}^\top\tilde{\mathbf{v}}$
for $t = 0$ to $K - 1$ **do**
  Define $Q_t(\mathbf{y}) = \frac{\mathbf{y}^\top B\mathbf{y}}{2} - \nabla W(\tilde{\mathbf{v}}_t)^\top\mathbf{y}$
  Let $\tilde{\varepsilon} = \frac{\lambda_{\min}(A)\varepsilon}{2}$
  Compute approximate minimizer $\hat{\mathbf{y}}_t$ of $Q_t(\mathbf{y})$ using Acc-SVRG, such that

$$\frac{1}{4}\|\hat{\mathbf{y}}_t - B^{-1}\nabla W(\tilde{\mathbf{v}}_t)\| \leq \min\left\{\frac{\tilde{\varepsilon}}{100(G_W+1)\|B\|^{1/2}}, 1\right\}$$

  $\tilde{\mathbf{v}}_{t+1} = \tilde{\mathbf{v}}_t - \frac{1}{4}\hat{\mathbf{y}}_t$
end for
Output $\tilde{\mathbf{v}}_K$ such that $\|A^{-1}\mathbf{b} - \tilde{\mathbf{v}}_K\| \leq \varepsilon$

Running Time Analysis: Define $h_t \triangleq W(\tilde{\mathbf{v}}_t) - \min_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}})$. Using the standard descent analysis, we show that the following holds true for $t \geq 0$:

$$h_t \leq \max\{\tilde{\varepsilon}, (0.9)^t h_0\}.$$

This follows directly from the (matrix norm-based) gradient descent analysis which we outline below. To make the analysis easier, we define a sequence of exact iterates as:

$$\mathbf{z}_{t+1} = \tilde{\mathbf{v}}_t - \frac{1}{4}B^{-1}\nabla W(\tilde{\mathbf{v}}_t).$$

Furthermore, our approximate solution $\hat{\mathbf{y}}_t$ is such that

$$\|\mathbf{z}_{t+1} - \tilde{\mathbf{v}}_{t+1}\| = \frac{1}{4}\|\hat{\mathbf{y}}_t - B^{-1}\nabla W(\tilde{\mathbf{v}}_t)\| \leq \min\left\{\frac{\tilde{\varepsilon}}{100(G_W+1)\|B\|^{1/2}}, 1\right\}, \qquad (2)$$

where G_W is a bound on $\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}}$. The bound G_W can be taken as a bound on the gradient of the quadratic at the start of the procedure (for $\tilde{\mathbf{v}}_0 = 0$), so it is enough to take $G_W = \|B^{-1}\|^{1/2}\|\mathbf{b}\|$, since $\|\nabla W(0)\|_{B^{-1}} \leq \|B^{-1}\|^{1/2}\|\nabla W(0)\| = \|B^{-1}\|^{1/2}\|\mathbf{b}\|$. We now have that

$$\begin{aligned}
h_{t+1} - h_t &= W(\tilde{\mathbf{v}}_{t+1}) - W(\tilde{\mathbf{v}}_t) \\
&\leq \langle \nabla W(\tilde{\mathbf{v}}_t), \tilde{\mathbf{v}}_{t+1} - \tilde{\mathbf{v}}_t \rangle + \|\tilde{\mathbf{v}}_{t+1} - \tilde{\mathbf{v}}_t\|_B^2 \\
&= \langle \nabla W(\tilde{\mathbf{v}}_t), \mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t \rangle + \langle \nabla W(\tilde{\mathbf{v}}_t), \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1} \rangle + \|\mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t + \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B^2 \\
&= \langle \nabla W(\tilde{\mathbf{v}}_t), \mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t \rangle + \langle \nabla W(\tilde{\mathbf{v}}_t), \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1} \rangle + \|\mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t\|_B^2 + \|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B^2 \\
&\qquad + 2\langle \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}, B(\mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t) \rangle \\
&= \langle \nabla W(\tilde{\mathbf{v}}_t), \mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t \rangle + \frac{1}{2}\langle \nabla W(\tilde{\mathbf{v}}_t), \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1} \rangle + \|\mathbf{z}_{t+1} - \tilde{\mathbf{v}}_t\|_B^2 + \|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B^2 \\
&\leq -\frac{1}{4}\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 + \frac{1}{2}\langle \nabla W(\tilde{\mathbf{v}}_t), \tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1} \rangle + \frac{1}{8}\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 + \|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B^2 \\
&\leq -\frac{1}{8}\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 + \frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}}\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B + \|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B^2 \\
&\leq -\frac{1}{8}\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 + \left(\frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}} + \|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B\right)\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B \\
&\leq -\frac{1}{8}\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 + \left(\frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}} + 1\right)\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B.
\end{aligned}$$

By $\frac{1}{2}$-strong convexity of $W(\cdot)$ w.r.t. $\|\cdot\|_B$, we have that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned}
W(\mathbf{y}) &\geq W(\mathbf{x}) + \nabla W(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{4}\|\mathbf{y} - \mathbf{x}\|_B^2 \\
&\geq \min_z \{W(\mathbf{x}) + \nabla W(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{4}\|\mathbf{y} - \mathbf{x}\|_B^2\} \\
&= W(\mathbf{x}) - \|\nabla W(\mathbf{x})\|_{B^{-1}}^2.
\end{aligned}$$

It follows that

$$-\|\nabla W(\tilde{\mathbf{v}}_t)\|_{B^{-1}}^2 \leq -h_t, \tag{3}$$

and so

$$h_{t+1} - h_t \leq -\frac{1}{8}h_t + \left(\frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}} + 1\right)\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B,$$

which gives us

$$\begin{aligned}
h_{t+1} &\leq 0.9h_t + \left(\frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}} + 1\right)\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\|_B \\
&\leq 0.9h_t + \left(\frac{1}{2}\|\nabla W(\tilde{\mathbf{v}})\|_{B^{-1}} + 1\right)\|B\|^{1/2}\|\tilde{\mathbf{v}}_{t+1} - \mathbf{z}_{t+1}\| \\
&\leq 0.9h_t + 0.01\tilde{\varepsilon},
\end{aligned}$$

where the final inequality follows by our approximation guarantee in (2).

Using the inductive assumption that $h_t \leq \max\{\tilde{\varepsilon}, (0.9)^t h_0\}$, it follows that

$$h_{t+1} \leq \max\{\tilde{\varepsilon}, (0.9)^{t+1} h_0\}.$$

Using the above inequality, it follows that for $t \geq O(\log(\frac{h_0}{\tilde{\varepsilon}}))$, we have that $h_t \leq \tilde{\varepsilon}$. Note that $W(\tilde{\mathbf{v}})$ is $\lambda_{\min}(A)$-strongly convex w.r.t. $\|\cdot\|$. Thus, we have that if $h_t \leq \tilde{\varepsilon}$, then

$$\frac{\lambda_{\min}(A)}{2}\|\tilde{\mathbf{v}}_t - \operatorname*{argmin}_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}})\| \leq h_t \leq \tilde{\varepsilon},$$

and so it follows that

$$\|\tilde{\mathbf{v}}_t - \operatorname*{argmin}_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}})\| \leq \frac{2\tilde{\varepsilon}}{\lambda_{\min}(A)}. \tag{4}$$

The running time of the above sub-procedure is bounded by the time to calculate $\nabla W(\tilde{\mathbf{v}})$, which takes at most $O(md)$ time, and the time required to compute $\hat{\mathbf{y}}_t$, which involves approximately solving a linear system in B at each step to $\hat{\varepsilon}$ accuracy, where

$$\hat{\varepsilon} \triangleq \min\left\{\frac{\tilde{\varepsilon}}{100(G_W+1)\|B\|^{1/2}}, 1\right\}.$$

Combining these we get that the total running time is

$$\tilde{O}(md + LIN(B,\hat{\varepsilon}))\log\left(\frac{1}{\tilde{\varepsilon}}\right).$$

Note that we set $\tilde{\varepsilon} = \frac{\lambda_{\min}(A)\varepsilon}{2}$, and so $\|\tilde{\mathbf{v}}_t - \operatorname{argmin}_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}})\| \leq \varepsilon$. Now we can bound $LIN(B,\hat{\varepsilon})$ by $\tilde{O}(d^2 + d\sqrt{\kappa(A)d})\log(1/\varepsilon)$ by using Acc-SVRG to solve the linear system and by noting that B is an $O(d\log(d))$ sized 2-approximation sample of A, which finishes the proof. ■

Acknowledgements

We would like to thank Ulysse Marteau-Ferey for pointing out this error to us.