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

Naman Agarwal
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

NAMANA@CS.PRINCETON.EDU

Brian Bullins
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

BBULLINS@CS.PRINCETON.EDU

Elad Hazan
Computer Science Department
Princeton University
Princeton, NJ 08540, USA

EHAZAN@CS.PRINCETON.EDU

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, $B = \sum_{i=1}^{O(d \log(d))} u_i u_i^T$ 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 $W(\tilde{v})$ (defined in Algorithm 4) is $\frac{1}{2}$-strongly convex and 2-smooth with respect to the norm given by $\|\tilde{v}\|_B \triangleq \sqrt{\tilde{v}^T B \tilde{v}}$. In this case, it is well-known that running an iterative method of the form

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

will converge to an $\varepsilon$-approximate minimizer of $W(\tilde{v})$ in $O(\log(h_0/\varepsilon))$ iterations, where $h_0 \triangleq W(\tilde{v}_0) - \min_{v} W(\tilde{v})$. Thus, all that is left is to handle the approximation error incurred by Acc-SVRG.
Algorithm 4: Fast Quadratic Solver (FQS) (Corrected)

1: Input: $A = \sum_{i=1}^{m} (v_i v_i^T + \lambda I)$, $b$, $\epsilon > 0$, $K = O(\log(1/\epsilon))$, $\tilde{v}_0 = 0$
2: Output: $\tilde{v}_K$ s.t. $\|A^{-1} b - \tilde{v}_K\| \leq \epsilon$
3: Compute $B$ s.t. $2B \succeq A \succeq \frac{1}{2} B$ using REPEATED HALVING (Algorithm 3)
4: Define $W(\tilde{v}) = \frac{1}{2} \tilde{v}^T A \tilde{v} - b^T \tilde{v}$
5: for $t = 0$ to $K - 1$ do
6: Define $Q_t(y) = y^T B y - \nabla W(\tilde{v}_t)^T y$
7: Let $\tilde{\epsilon} = \frac{\lambda_{\min}(A)\epsilon}{2}$
8: Compute approximate minimizer $\hat{y}_t$ of $Q_t(y)$ using Acc-SVRG, such that
9: $\tilde{v}_{t+1} = \tilde{v}_t - \frac{1}{4} \hat{y}_t$
10: end for
11: Output $\tilde{v}_K$ such that $\|A^{-1} b - \tilde{v}_K\| \leq \epsilon$

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

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

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:

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

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

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

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

\[ h_{t+1} - h_t = W(\tilde{v}_{t+1}) - W(\tilde{v}_t) \]
\[ \leq \langle \nabla W(\tilde{v}_t), \tilde{v}_{t+1} - \tilde{v}_t \rangle + \| \tilde{v}_{t+1} - \tilde{v}_t \|_B^2 \]
\[ = \langle \nabla W(\tilde{v}_t), z_{t+1} - \tilde{v}_t \rangle + \langle \nabla W(\tilde{v}_t), \tilde{v}_{t+1} - z_{t+1} \rangle + \| z_{t+1} - \tilde{v}_t + \tilde{v}_{t+1} - z_{t+1} \|_B^2 \]
\[ = \langle \nabla W(\tilde{v}_t), z_{t+1} - \tilde{v}_t \rangle + \langle \nabla W(\tilde{v}_t), \tilde{v}_{t+1} - z_{t+1} \rangle + \| z_{t+1} - \tilde{v}_t \|_B^2 + \| \tilde{v}_{t+1} - z_{t+1} \|_B^2 \]
\[ + 2 \langle \tilde{v}_{t+1} - z_{t+1}, B(z_{t+1} - \tilde{v}_t) \rangle \]
\[ = \langle \nabla W(\tilde{v}_t), z_{t+1} - \tilde{v}_t \rangle + \frac{1}{2} \langle \nabla W(\tilde{v}_t), \tilde{v}_{t+1} - z_{t+1} \rangle + \| z_{t+1} - \tilde{v}_t \|_B^2 + \| \tilde{v}_{t+1} - z_{t+1} \|_B^2 \]
\[ \leq - \frac{1}{4} \| \nabla W(\tilde{v}_t) \|_{B^{-1}}^2 + \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} \| \tilde{v}_{t+1} - z_{t+1} \|_B + \| \tilde{v}_{t+1} - z_{t+1} \|_B^2 \]
\[ \leq - \frac{1}{8} \| \nabla W(\tilde{v}_t) \|_{B^{-1}}^2 + \left( \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} + \| \tilde{v}_{t+1} - z_{t+1} \|_B \right) \| \tilde{v}_{t+1} - z_{t+1} \|_B \]
\[ \leq - \frac{1}{8} \| \nabla W(\tilde{v}_t) \|_{B^{-1}}^2 + \left( \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} + 1 \right) \| \tilde{v}_{t+1} - z_{t+1} \|_B. \]

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

\[ W(y) \geq W(x) + \nabla W(x) \top (y - x) + \frac{1}{4} \| y - x \|_B^2 \]
\[ \geq \min \{ W(x) + \nabla W(x) \top (y - x) + \frac{1}{4} \| y - x \|_B^2 \} \]
\[ = W(x) - \| \nabla W(x) \|_{B^{-1}}^2. \]

It follows that

\[ - \| \nabla W(\tilde{v}_t) \|_{B^{-1}}^2 \leq -h_t, \]

and so

\[ h_{t+1} - h_t \leq - \frac{1}{8} h_t + \left( \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} + 1 \right) \| \tilde{v}_{t+1} - z_{t+1} \|_B, \]

which gives us

\[ h_{t+1} \leq 0.9 h_t + \left( \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} + 1 \right) \| \tilde{v}_{t+1} - z_{t+1} \|_B \]
\[ \leq 0.9 h_t + \left( \frac{1}{2} \| \nabla W(\tilde{v}_t) \|_{B^{-1}} + 1 \right) \| B \|^{1/2} \| \tilde{v}_{t+1} - z_{t+1} \| \]
\[ \leq 0.9 h_t + 0.01 \bar{\epsilon}, \]

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

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

\[ h_{t+1} \leq \max \{ \bar{\epsilon}, (0.9)^{t+1} h_0 \}. \]
Using the above inequality, it follows that for \( t \geq O(\log(h_{t0}^2)) \), we have that \( h_t \leq \tilde{\epsilon} \). Note that \( W(\tilde{v}) \) is \( \lambda_{\text{min}}(A) \)-strongly convex w.r.t. \( \|\cdot\| \). Thus, we have that if \( h_t \leq \tilde{\epsilon} \), then

\[
\frac{\lambda_{\text{min}}(A)}{2} \| \tilde{v}_t - \arg\min_{\tilde{v}} W(\tilde{v}) \| \leq h_t \leq \tilde{\epsilon},
\]

and so it follows that

\[
\| \tilde{v}_t - \arg\min_{\tilde{v}} W(\tilde{v}) \| \leq \frac{2\tilde{\epsilon}}{\lambda_{\text{min}}(A)}.
\]  

(4)

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

\[
\hat{\epsilon} \triangleq \min \left\{ \frac{\epsilon}{100(G_W + 1)\|B\|^{1/2}}, 1 \right\}.
\]

Combining these we get that the total running time is

\[
\tilde{O}(md + LIN(B, \hat{\epsilon})) \log \left( \frac{1}{\epsilon} \right).
\]

Note that we set \( \tilde{\epsilon} = \frac{\lambda_{\text{min}}(A)\epsilon}{2} \), and so \( \| \tilde{v}_t - \arg\min_{\tilde{v}} W(\tilde{v}) \| \leq \epsilon \). Now we can bound \( LIN(B, \hat{\epsilon}) \) by \( \tilde{O}(d^2 + d\sqrt{\kappa(A)d}) \log(1/\epsilon) \) 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.

\[\Box\]

Acknowledgements

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