Appendices: Proofs of Lemmas and Theorems in "Adaptive Stochastic Variance Reduction for Subsampled Newton Method with Cubic Regularization"

Junyu Zhang* Lin Xiao† Shuzhong Zhang‡

September 6, 2020

1 Proof of Lemma 2.2

Lemma 2.2 *Let $Z_1, ..., Z_n$ be i.i.d. random matrices in $\mathbb{R}^{d\times d}$ with $\mathbb{E}[Z_1] = 0$ and $\mathbb{E}\left[\|Z_1\|_F^4\right] < \infty$. Then*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_F^4\right] \le \frac{3}{n^2}\mathbb{E}\left[\|Z_1\|_F^4\right].$$

Proof. Let $\langle Z_1, Z_2\rangle = \mathrm{trace}(Z_1^T Z_2)$ be the inner product of two matrices. Then $\|Z_1\|_F^2 = \langle Z_1, Z_1\rangle$. Using the assumption that $\mathbb{E}[Z_i] = 0$ for all i and $Z_1, \ldots, Z_n$ are independent and identically distributed, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_F^4\right] = \frac{1}{n^4}\mathbb{E}\left[\left\langle\sum_{i=1}^n Z_i, \sum_{i=1}^n Z_i\right\rangle^2\right] = \frac{1}{n^4}(T_1 + T_2 + \cdots + T_7), \tag{1}$$

where

$$\begin{aligned}
T_1 &= n\mathbb{E}\left[\|Z_1\|_F^4\right], \\
T_2 &= 4n(n-1)\mathbb{E}\left[\|Z_1\|_F^2\langle Z_1, Z_2\rangle\right] = 0, \\
T_3 &= 2n(n-1)\mathbb{E}\left[\langle Z_1, Z_2\rangle^2\right] \le 2n(n-1)\mathbb{E}\left[\|Z_1\|_F^4\right], \\
T_4 &= n(n-1)\mathbb{E}\left[\|Z_1\|_F^2\|Z_2\|_F^2\right] \le n(n-1)\mathbb{E}\left[\|Z_1\|_F^4\right], \\
T_5 &= 4n(n-1)(n-2)\mathbb{E}\left[\langle Z_1, Z_2\rangle\langle Z_2, Z_3\rangle\right] = 0, \\
T_6 &= 2n(n-1)(n-2)\mathbb{E}\left[\langle Z_1, Z_2\rangle\|Z_3\|_F^2\right] = 0, \\
T_7 &= n(n-1)(n-2)(n-3)\mathbb{E}\left[\langle Z_1, Z_2\rangle\langle Z_3, Z_4\rangle\right] = 0.
\end{aligned}$$

In total, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_F^4\right] = \frac{3n^2 - 2n}{n^4}\mathbb{E}\left[\|Z_1\|_F^4\right] \le \frac{3}{n^2}\mathbb{E}\left[\|Z_1\|_F^4\right].$$

which is the desired result. □

*Department of Industrial and System Engineering, University of Minnesota (zhan4393@umn.edu).
†Machine Learning and Optimization Group, Microsoft Research, Redmond, WA (lin.xiao@microsoft.com).
‡Department of Industrial and System Engineering, University of Minnesota (zhangs@umn.edu).

2 Proof of Lemma 2.3

Lemma 2.3 *Let the variance reduced gradient g_t^k and Hessian H_t^k be constructed according to*

$$g_t^k = \frac{1}{|\mathcal{S}_t^k|}\sum_{i\in\mathcal{S}_t^k}\Big(\nabla f_i(x_t^k) - \nabla f_i(\tilde{x}^{k-1})\Big) + \tilde{g}^{k-1},$$
$$H_t^k = \frac{1}{|\mathcal{B}_t^k|}\sum_{i\in\mathcal{B}_t^k}\Big(\nabla^2 f_i(x_t^k) - \nabla^2 f_i(\tilde{x}^{k-1})\Big) + \tilde{H}^{k-1},$$

with $\tilde{g}^{k-1} = \nabla F(\tilde{x}^{k-1})$ and $\tilde{H}^{k-1} = \nabla^2 F(\tilde{x}^{k-1})$. Then they satisfy the following equalities and inequalities

$$\mathbb{E}\big[H_t^k \,\big|\, x_t^k\big] = \nabla^2 F(x_t^k),$$
$$\mathbb{E}\left[\big\|H_t^k - \nabla^2 F(x_t^k)\big\|_F^2 \,\big|\, x_t^k\right] \le \frac{\rho^2}{|\mathcal{B}_t^k|}\big\|x_t^k - \tilde{x}^{k-1}\big\|^2,$$
$$\mathbb{E}\left[\big\|H_t^k - \nabla^2 F(x_t^k)\big\|_F^4 \,\big|\, x_t^k\right] \le \frac{33\rho^4}{|\mathcal{B}_t^k|^2}\big\|x_t^k - \tilde{x}^{k-1}\big\|^4,$$

and

$$\mathbb{E}\left[g_t^k \,\big|\, x_t^k\right] = \nabla F(x_t^k),$$
$$\mathbb{E}\left[\big\|g_t^k - \nabla F(x_t^k)\big\|_F^2 \,\big|\, x_t^k\right] \le \frac{L^2}{|\mathcal{S}_t^k|}\big\|x_t^k - \tilde{x}^{k-1}\big\|^2.$$

Proof. First, we prove the bounds for the variance reduced Hessian estimate. It is straightforward to show that $\mathbb{E}\left[H_t^k | x_t^k\right] = \nabla^2 F(x_t^k)$. For the rest two inequalities, let us first define

$$Z_j = \nabla^2 f_j(x_t^k) - \nabla^2 f_j(\tilde{x}^{k-1}) + \nabla^2 F(\tilde{x}^{k-1}) - \nabla^2 F(x_t^k),$$

where j is a uniform sample from $\{1, ..., N\}$. Note that

$$\mathbb{E}\left[\nabla^2 f_j(x_t^k) - \nabla^2 f_j(\tilde{x}^{k-1})\,\Big|\, x_t^k\right] = \nabla^2 F(x_t^k) - \nabla^2 F(\tilde{x}^{k-1}).$$

Therefore, $\mathbb{E}\left[Z_j | x_t^k\right] = 0$. For the ease of notation, define $z_j = \nabla^2 f_j(x_t^k) - \nabla^2 f_j(\tilde{x}^{k-1})$ such that $Z_j = z_j - \mathbb{E}\big[z_j | x_t^k\big]$. According to the Lipschitz continuity conditions, $\|z_j\|_F \le \rho\|x_t^k - \tilde{x}^{k-1}\|$. For the second moment, we have

$$\begin{aligned}
\mathbb{E}\left[\|Z_j\|_F^2 \Big| x_t^k\right] &= \mathbb{E}\left[\big\|z_j - \mathbb{E}[z_j | x_t^k]\big\|_F^2 \Big| x_t^k\right] \\
&= \mathbb{E}\left[\|z_j\|_F^2 \big| x_t^k\right] - \big\|\mathbb{E}[z_j | x_t^k]\big\|_F^2 \\
&\le \mathbb{E}\left[\|z_j\|_F^2 \big| x_t^k\right] \\
&\le \rho^2\big\|x_t^k - \tilde{x}^{k-1}\big\|^2.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^2 \middle| x_t^k\right] = \mathbb{E}\left[\left\|\frac{1}{|\mathcal{B}_t^k|}\sum_{j\in\mathcal{B}_t^k} Z_j\right\|_F^2 \middle| x_t^k\right] = \frac{1}{|\mathcal{B}_t^k|}\mathbb{E}\left[\|Z_1\|_F^2 | x_t^k\right] \le \frac{\rho^2}{|\mathcal{B}_t^k|}\left\|x_t^k - \tilde{x}^{k-1}\right\|^2.$$

For the fourth moment, we have

$$\begin{aligned}
\mathbb{E}\left[\|Z_j\|_F^4 | x_t^k\right] &= \mathbb{E}\left[\left\|z_j - \mathbb{E}[z_j|x_t^k]\right\|_F^4 | x_t^k\right] \\
&= \mathbb{E}\left[\|z_j\|_F^4 | x_t^k\right] + 4\mathbb{E}\left[\left\langle z_j, \mathbb{E}\left[z_j|x_t^k\right]\right\rangle^2 | x_t^k\right] + 2\mathbb{E}\left[\|z_j\|_F^2 | x_t^k\right]\left\|\mathbb{E}[z_j|x_t^k]\right\|_F^2 \\
&\quad - 4\left\langle \mathbb{E}\left[\|z_j\|_F^2 z_j | x_t^k\right], \mathbb{E}[z_j|x_t^k]\right\rangle - 3\left\|\mathbb{E}[z_j|x_t^k]\right\|_F^4 \\
&\le 11\,\mathbb{E}\left[\|z_j\|_F^4 | x_t^k\right] \\
&\le 11\rho^4\left\|x_t^k - \tilde{x}^{k-1}\right\|^4,
\end{aligned} \tag{2}$$

where we used $\langle z_j, \mathbb{E}[z_j|x_t^k]\rangle \le \|z_j\|_F\left\|\mathbb{E}[z_j|x_t^k]\right\|_F$ and $\left\|\mathbb{E}[z_j|x_t^k]\right\|_F^2 \le \mathbb{E}\left[\|z_j\|_F^2|x_t^k\right]$. Then by Lemma 2.2, we obtain

$$\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^4 \middle| x_t^k\right] = \mathbb{E}\left[\left\|\frac{1}{|\mathcal{B}_t^k|}\sum_{j\in\mathcal{B}_t^k} Z_j\right\|_F^4 \middle| x_t^k\right] \le \frac{3}{|\mathcal{B}_t^k|^2}\mathbb{E}\left[\|Z_1\|_F^4 | x_t^k\right] \le \frac{33\rho^4}{|\mathcal{B}_t^k|^2}\left\|x_t^k - \tilde{x}^{k-1}\right\|^4.$$

Similarly, one can get the variance bounds for the gradient estimate. This part of the proof is omitted to avoid repetition. □

3 Proof of Corollary 2.4

Corollary 2.4 *Let H_t^k, g_t^k, ξ_t^k and the mini-batch index sets $\mathcal{B}_t^k$ and $\mathcal{S}_t^k$ be generated according to Algorithm 1. Then we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F \,|\, x_t^k\right] &\le \rho\left(\|\xi_{t-1}^k\| + \epsilon^{1/2}\right), \\
\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^2 \,|\, x_t^k\right] &\le \rho^2\left(\|\xi_{t-1}^k\|^2 + \epsilon\right), \\
\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^3 \,|\, x_t^k\right] &\le 33^{3/4}\rho^3\left(\|\xi_{t-1}^k\|^3 + \epsilon^{3/2}\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\left\|g_t^k - \nabla F(x_t^k)\right\|_F \,|\, x_t^k\right] &\le L\left(\|\xi_{t-1}^k\|^2 + \epsilon\right), \\
\mathbb{E}\left[\left\|g_t^k - \nabla F(x_t^k)\right\|_F^{3/2} \,|\, x_t^k\right] &\le L^{3/2}\left(\|\xi_{t-1}^k\|^3 + \epsilon^{3/2}\right).
\end{aligned}$$

Proof. By Lemma 2.3 and the mini-batch size rule in Algorithm 1,

$$\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^2 | x_t^k\right] \;\le\; \frac{\rho^2}{|\mathcal{B}_t^k|}\left\|x_t^k - \tilde{x}^{k-1}\right\|^2 \;\le\; \rho^2\epsilon_H \;\le\; \rho^2\left(\|\xi_{t-1}^k\|^2 + \epsilon\right).$$

Due to the concavity of the square root function $\sqrt{\cdot}$, we can apply Jensen's inequality to obtain

$$\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F \middle| x_t^k\right] \;\le\; \sqrt{\mathbb{E}\left[\|H_t^k - \nabla^2 F(x_t^k)\|_F^2 \middle| x_t^k\right]} \;\le\; \rho\sqrt{\epsilon_H} \;\le\; \rho\left(\|\xi_{t-1}^k\| + \epsilon^{1/2}\right).$$

Similarly, we have

$$\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^4 \middle| x_t^k\right] \;\le\; \frac{33\rho^4}{|\mathcal{B}_t^k|^2}\left\|x_t^k - \tilde{x}^{k-1}\right\|^4 \;\le\; 33\rho^4\epsilon_H^2.$$

Again, with the concavity of the function $(\cdot)^{3/4}$, applying Jensen's inequality yields

$$\begin{aligned}\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^3 \middle| x_t^k\right] \le \left(\mathbb{E}\left[\left\|H_t^k - \nabla^2 F(x_t^k)\right\|_F^4 \middle| x_t^k\right]\right)^{3/4} \;&\le\; 33^{3/4}\rho^3\epsilon_H^{3/2} \\ &\le\; 33^{3/4}\rho^3\left(\|\xi_{t-1}^k\|^3 + \epsilon^{3/2}\right).\end{aligned}$$

Following a similar line of arguments, one can get the bounds for the gradient variances. We omit the details to avoid repetition. □

4 Proof of Lemma 2.10

Lemma 2.10 *Suppose $X_1, \ldots, X_N$ are matrices in $\mathbb{R}^{d\times d}$ satisfying $\frac{1}{N}\sum_{i=1}^N X_i = 0$. Let $Z_1, ..., Z_n$, where $n \le N$, be uniformly sampled from $X_1, ..., X_N$ without replacement. Then*

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_F^2\right] &= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\mathbb{E}\left[\|Z_1\|_F^2\right], \\ \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_F^4\right] &= \frac{1}{n^4}\left(r_1\mathbb{E}\left[\|Z_1\|_F^4\right] + r_2\mathbb{E}\left[\left\langle Z_1, Z_2\right\rangle^2\right] + r_3\mathbb{E}\left[\|Z_1\|_F^2\|Z_2\|_F^2\right]\right),\end{aligned}$$

where $\langle Z_i, Z_j\rangle = \mathrm{trace}\big(Z_i^T Z_j\big)$ and

$$\begin{aligned}r_1 &= n\left(1 - 4\cdot\frac{n-1}{N-1} + 6\cdot\frac{(n-1)(n-2)}{(N-1)(N-2)} - 3\cdot\frac{(n-1)(n-2)(n-3)}{(N-1)(N-2)(N-3)}\right), \\ r_2 &= n(n-1)\left(2 - 4\cdot\frac{n-2}{N-2} + 2\cdot\frac{(n-2)(n-3)}{(N-2)(N-3)}\right), \\ r_3 &= n(n-1)\left(1 - 2\cdot\frac{n-2}{N-2} + 1\cdot\frac{(n-2)(n-3)}{(N-2)(N-3)}\right).\end{aligned}$$

Proof. The first equation is a standard result for variance analysis of sampling without replacement scheme. Hence we omit the proof of this equation. To prove the second equation, we start with the expansions in (1) and calculate the terms $T_1, ..., T_7$ for sampling without replacement. First, the following three terms do not change:

$$\begin{aligned}T_1 &= n\mathbb{E}\left[\|Z_1\|_F^4\right], \\ T_3 &= 2n(n-1)\mathbb{E}\left[\left(\langle Z_1, Z_2\rangle\right)^2\right], \\ T_4 &= n(n-1)\mathbb{E}\left[\|Z_1\|_F^2\|Z_2\|_F^2\right].\end{aligned}$$

For T_2, we have

$$\begin{aligned}
T_2 &= 4n(n-1)\mathbb{E}\left[\|Z_1\|_F^2\langle Z_1, Z_2\rangle\right] \\
&= 4n(n-1)\mathbb{E}\Big[\|Z_1\|_F^2\big\langle Z_1, \mathbb{E}\left[Z_2|Z_1\right]\big\rangle\Big] \\
&= 4n(n-1)\sum_{i=1}^{N}\frac{1}{N}\|X_i\|_F^2\Big\langle X_i, \sum_{j\neq i}\frac{1}{N-1}X_j\Big\rangle \\
&= \frac{4n(n-1)}{(N)(N-1)}\sum_{i=1}^{N}\|X_i\|_F^2\Big\langle X_i, \underbrace{\sum_{j=1}^{N}X_j}_{0} - X_i\Big\rangle \\
&= -\frac{4n(n-1)}{N-1}\sum_{i=1}\frac{1}{N}\|X_i\|_F^4 \\
&= -\frac{4n(n-1)}{N-1}\mathbb{E}\left[\|Z_1\|_F^4\right].
\end{aligned}$$

For T_5, we have

$$\begin{aligned}
T_5 &= 4n(n-1)(n-2)\mathbb{E}\big[\langle Z_1, Z_2\rangle\cdot\langle Z_2, Z_3\rangle\big] \\
&= 4n(n-1)(n-2)\mathbb{E}\Big[\big\langle\mathbb{E}\big[Z_1|Z_2\big], Z_2\big\rangle\cdot\big\langle Z_2, \mathbb{E}\big[Z_3|Z_2, Z_1\big]\big\rangle\Big] \\
&= \frac{4n(n-1)(n-2)}{N(N-1)(N-2)}\sum_{i=1}^{N}\Big\langle X_i, \sum_{j\neq i}X_j\Big(\Big\langle X_i, \sum_{k\neq i,j}X_k\Big\rangle\Big)\Big\rangle \\
&= \frac{4n(n-1)(n-2)}{N(N-1)(N-2)}\sum_{i=1}^{N}\Big\langle X_i, \sum_{j\neq i}X_j\Big(\big\langle X_i, -X_i - X_j\big\rangle\Big)\Big\rangle \\
&= \frac{4n(n-1)(n-2)}{N(N-1)(N-2)}\sum_{i=1}^{N}\Big\langle X_i, \sum_{j\neq i}X_j\Big(-\|X_i\|_F^2 - \langle X_i, X_j\rangle\Big)\Big\rangle \\
&= -\frac{4n(n-1)(n-2)}{N-2}\sum_{i=1}^{N}\sum_{j\neq i}\frac{1}{N(N-1)}\Big(\|X_i\|_F^2\langle X_i, X_j\rangle + \big(\langle X_i, X_j\rangle\big)^2\Big) \\
&= -\frac{4n(n-1)(n-2)}{N-2}\Big(\mathbb{E}\left[\|Z_1\|_F^2\langle Z_1, Z_2\rangle\right] + \mathbb{E}\left[\big(\langle Z_1, Z_2\rangle\big)^2\right]\Big) \\
&= 4n\frac{(n-1)(n-2)}{(N-1)(N-2)}\mathbb{E}\left[\|Z_1\|_F^4\right] - \frac{4n(n-1)(n-2)}{N-2}\mathbb{E}\left[\big(\langle Z_1, Z_2\rangle\big)^2\right],
\end{aligned}$$

where in the last equality we used result for T_2. In similar ways, one can find the expressions for T_6 and T_7:

$$\begin{aligned}
T_6 &= 2n(n-1)(n-2)\mathbb{E}\left[\|Z_3\|_F^2\langle Z_1, Z_2\rangle\right] \\
&= 2n\frac{(n-1)(n-2)}{(N-1)(N-2)}\mathbb{E}\left[\|Z_1\|_F^4\right] - 2n(n-1)\frac{n-2}{N-2}\mathbb{E}\left[\|Z_1\|_F^2\|Z_2\|_F^2\right]
\end{aligned}$$

and

$$\begin{aligned} T_7 &= 2n(n-1)(n-2)(n-3)\mathbb{E}\big[\langle Z_1, Z_2\rangle \cdot \langle Z_3, Z_4\rangle\big] \\ &= -3n\frac{(n-1)(n-1)(n-3)}{(N-1)(N-2)(N-3)}\mathbb{E}\left[\|Z_1\|_F^4\right] + 2n(n-1)\frac{(n-2)(n-3)}{(N-2)(N-3)}\mathbb{E}\big[\big(\langle Z_1, Z_2\rangle\big)^2\big] \\ &\quad +n(n-1)\frac{(n-2)(n-3)}{(N-2)(N-3)}\mathbb{E}\big[\|Z_1\|_F^2\|Z_2\|_F^2\big]. \end{aligned}$$

Summing these terms up gives the desired result. □

5 Proof of Lemma 3.1

Lemma 3.1 *For any $a, b, \theta_1, \theta_2 > 0$, the following inequality holds:*

$$(a+b)^3 \le \left(1 + 2\theta_1^{-3} + \theta_2^{-6}\right) a^3 + \left(1 + \theta_1^6 + 2\theta_2^3\right) b^3.$$

Proof. We expand $(a+b)^3$ and then use Young's inequality,

$$\begin{aligned} (a+b)^3 &= a^3 + b^3 + 3a^2b + 3b^2a \\ &= a^3 + b^3 + 3(a/\theta_1)^2(b\theta_1^2) + 3(a/\theta_2^2)(b\theta_2)^2 \\ &\le a^3 + b^3 + 3\left(\frac{(a^2/\theta_1^2)^{3/2}}{3/2} + \frac{(b\theta_1^2)^3}{3}\right) + 3\left(\frac{(a/\theta_2^2)^3}{3} + \frac{(b^2\theta_2^2)^{3/2}}{3/2}\right) \\ &= (1 + 2\theta_1^{-3} + \theta_2^{-6})a^3 + (1 + \theta_1^6 + 2\theta_2^3)b^3. \end{aligned}$$

This completes the proof. □