Tracking the gradients using the Hessian: A new look at variance reducing stochastic methods
Appendix

February 14, 2018

A Theoretical Analysis: Proofs

Following Bubeck (2015), we consider a single epoch of SVRG and its extensions, that is, $\bar{\theta} \in \Theta$, and the iteration, started from $\theta_0 = \bar{\theta}$:

$$\theta_t = \Pi_\Theta\Big(\theta_{t-1} - \gamma\Big[f'_{i_t}(\theta_{t-1}) - z_{i_t}(\theta_{t-1}) + \frac{1}{N}\sum_{j=1}^{N} z_j(\theta_{t-1})\Big]\Big),$$

with i_t uniformly at random in $\{1, \ldots, N\}$.

We also recall that, while the results are given using R the radius of the data, they can be readily transposed to $L_{\max}$ using $L_{\max} = R^2$.

We have, with $\mathcal{F}_{t-1}$ representing the information up to time t:

$$\begin{aligned}
\mathbf{E}\left[\|\theta_t - \theta_*\|^2 | \mathcal{F}_{t-1}\right] &\leqslant \mathbf{E}\left[\Big\|\theta_{t-1} - \theta_* - \gamma\Big[f'_{i_t}(\theta_{t-1}) - z_{i_t}(\theta_{t-1}) + \frac{1}{N}\sum_{j=1}^{N} z_j(\theta_{t-1})\Big]\Big\|^2 | \mathcal{F}_{t-1}\right] \\
&\qquad \text{by contractivity of projections,} \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 - 2\gamma F'(\theta_{t-1})^\top(\theta_{t-1} - \theta_*) + \gamma^2\|F'(\theta_{t-1})\|^2 \\
&\quad + \frac{\gamma^2}{N}\sum_{i=1}^{N}\Big\|f'_i(\theta_{t-1}) - z_i(\theta_{t-1}) - \frac{1}{N}\sum_{j=1}^{N} f'_j(\theta_{t-1}) + \frac{1}{N}\sum_{j=1}^{N} z_j(\theta_{t-1})\Big\|^2 \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 - 2\gamma F'(\theta_{t-1})^\top(\theta_{t-1} - \theta_*) + \gamma^2\|F'(\theta_{t-1})\|^2 \\
&\quad + \frac{\gamma^2}{N}\sum_{i=1}^{N}\big\|f'_i(\theta_{t-1}) - z_i(\theta_{t-1})\big\|^2, \text{ by bounding the variance by the second moment.}
\end{aligned}$$

In the following sections, we provide proofs for several algorithms we consider in this paper.

A.1 SVRG

For regular SVRG (we provide the proof for completeness and because we need it later), we have: $z_i(\theta) = f_i'(\bar\theta)$ and we consider the bound

$$\begin{aligned}
\frac{\gamma^2}{N}\sum_{i=1}^N \big\|f_i'(\theta_{t-1}) - z_i(\theta_{t-1})\big\|^2 &\leqslant \frac{2\gamma^2}{N}\sum_{i=1}^N \big\|f_i'(\theta_{t-1}) - f_i'(\theta_*)\big\|^2 + \frac{2\gamma^2}{N}\sum_{i=1}^N \big\|f_i'(\bar\theta) - f_i'(\theta_*)\big\|^2 \\
&\leqslant 2\gamma^2 R^2 F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + 2\gamma^2R^2[F(\bar\theta) - F(\theta_*)]
\end{aligned}$$

leading to

$$\begin{aligned}
\mathbf{E}\big[\|\theta_t - \theta_*\|^2 | \mathcal{F}_{t-1}\big] &\leqslant \|\theta_{t-1}-\theta_*\|^2 - 2\gamma F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + \gamma^2\|F'(\theta_{t-1})\|^2 \\
&\quad +2\gamma^2R^2F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + 2\gamma^2R^2[F(\bar\theta) - F(\theta_*)].
\end{aligned}$$

Thus if $\gamma \leqslant 1/(2R^2+L)$, we get

$$\mathbf{E}\big[\|\theta_t - \theta_*\|^2 | \mathcal{F}_{t-1}\big] \leqslant \|\theta_{t-1}-\theta_*\|^2 - \gamma[F(\bar\theta_{t-1}) - F(\theta_*)] + 2\gamma^2R^2[F(\bar\theta) - F(\theta_*)].$$

This implies that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbf{E}\big[F(\bar\theta_{t-1}) - F(\theta_*)\big] &\leqslant \frac{1}{\gamma T}\|\bar\theta - \theta_*\|^2 + 2\gamma R^2[F(\bar\theta) - F(\theta_*)] \\
\mathbf{E}\bigg[F\Big(\frac{1}{T}\sum_{t=1}^T\bar\theta_{t-1}\Big) - F(\theta_*)\bigg] &\leqslant \Big(\frac{2}{\mu\gamma T} + 2\gamma R^2\Big)[F(\bar\theta) - F(\theta_*)].
\end{aligned}$$

This implies that if $\gamma = \frac{1}{4R^2}$ and $T \geqslant 8/(\gamma\mu) = \frac{32R^2}{\mu}$, then

$$\mathbf{E}\bigg[F\Big(\frac{1}{T}\sum_{t=1}^T\bar\theta_{t-1}\Big) - F(\theta_*)\bigg] \leqslant \frac{3}{4}[F(\bar\theta) - F(\theta_*)].$$

Thus, after $K = O(\log\frac{1}{\varepsilon})$ epochs of SVRG we have attained the required precision, which makes an overall access to gradients of $KN + KT = \big(N + \frac{R^2}{\mu}\big)\log\frac{1}{\varepsilon}$.

A.2 SVRG-2

We assume that $\frac{4\beta^2R^4}{\alpha}D^2 \leqslant L$ and $\gamma = 1/(4L)$.

In this situation, with no approximation, we have $z_i(\theta) = f_i'(\bar\theta) + f_i''(\bar\theta)(\theta - \bar\theta)$

and:

$$
\begin{aligned}
& \frac{\gamma^2}{N} \sum_{i=1}^N \big\| f_i'(\theta_{t-1}) - z_i(\theta_{t-1}) \big\|^2 \\
\leqslant \; & \frac{\gamma^2}{N} \sum_{i=1}^N R^2 \big[\frac{\beta}{2} (x_i^\top \theta_{t-1} - x_i^\top \bar\theta)^2 \big]^2 = \frac{\gamma^2 \beta^2 R^2}{4N} \sum_{i=1}^N \, (x_i^\top (\theta_{t-1} - \bar\theta))^4 \text{ using the bound on } \varphi''', \\
\leqslant \; & \frac{\gamma^2 \beta^2 R^2}{N} \sum_{i=1}^N \big[2(x_i^\top (\theta_{t-1} - \theta_*))^4 + 2(x_i^\top (\theta_* - \bar\theta))^4 \big] \\
\leqslant \; & \frac{\gamma^2 \beta^2 R^2}{N} \sum_{i=1}^N \big[2R^2 \|\theta_{t-1} - \theta_*\|^2 (x_i^\top (\theta_{t-1} - \theta_*))^2 + 2R^2 \|\bar\theta - \theta_*\|^2 (x_i^\top (\bar\theta - \theta_*))^2 \big] \text{ using } \|x_i\| \leqslant R, \\
\leqslant \; & \frac{2\gamma^2 \beta^2 R^4}{N} \|\theta_{t-1} - \theta_*\|^2 \sum_{i=1}^N (x_i^\top (\theta_{t-1} - \theta_*))^2 + \frac{2\gamma^2 \beta^2 R^4}{N} \|\bar\theta - \theta_*\|^2 \sum_{i=1}^N (x_i^\top (\bar\theta - \theta_*))^2 \\
\leqslant \; & \frac{4\gamma^2 \beta^2 R^4}{\alpha} \|\theta_{t-1} - \theta_*\|^2 [F(\theta_{t-1}) - F(\theta_*)] + \frac{4\gamma^2 \beta^2 R^4}{\alpha} \|\bar\theta - \theta_*\|^2 [F(\bar\theta) - F(\theta_*)] \text{ using } \varphi'' \geqslant \alpha, \\
\leqslant \; & \frac{4\gamma^2 \beta^2 R^4}{\alpha} D^2 [F(\theta_{t-1}) - F(\theta_*)] + \frac{4\gamma^2 \beta^2 R^4}{\alpha} D^2 [F(\bar\theta) - F(\theta_*)], \text{ using the compactness of } \Theta.
\end{aligned}
$$

With our assumptions, we have $\gamma \bigg(L + \frac{4\beta^2 R^4}{\alpha} D^2 \bigg) \leqslant 1$, and we get that

$$
\begin{aligned}
\mathbf{E}\left[\|\theta_t - \theta_*\|^2 | \mathcal{F}_{t-1} \right] & \leqslant \|\theta_{t-1} - \theta_*\|^2 - \gamma F'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*) + \frac{4\gamma^2 \beta^2 R^4}{\alpha} D^2 [F(\bar\theta) - F(\theta_*)] \\
& \leqslant \|\theta_{t-1} - \theta_*\|^2 - \gamma [F(\theta_{t-1}) - F(\theta_*)] + \frac{4\gamma^2 \beta^2 R^4}{\alpha} D^2 [F(\bar\theta) - F(\theta_*)].
\end{aligned}
$$

This leads to, with $\boxed{T \geqslant 4/(\mu\gamma) = \frac{16L}{\mu}}$ and using $\gamma \bigg(\frac{4\beta^2 R^4}{\alpha} D^2 \bigg) \leqslant 1/2$,

$$
\begin{aligned}
\mathbf{E}\left[F\Big(\frac{1}{T} \sum_{t=1}^T \bar\theta_{t-1} \Big) - F(\theta_*) \right] & \leqslant \bigg(\frac{2}{\mu\gamma T} + \frac{4\gamma\beta^2 R^4}{\alpha} D^2 \bigg) [F(\bar\theta) - F(\theta_*)] \\
& \leqslant \frac{3}{4} [F(\bar\theta) - F(\theta_*)].
\end{aligned}
$$

Thus, after $K = O(\log \frac{1}{\varepsilon})$ epochs of SVRG we have attained the required precision, which makes an overall access to gradients of $KN + KT = \big(N + \frac{L}{\mu} \big) \log \frac{1}{\varepsilon}$.

A.3 Stability of SVRG-2

If we make no compactness assumption on Θ, then we have:

$$
\begin{aligned}
& \frac{\gamma^2}{N}\sum_{i=1}^{N}\big\|f_i'(\theta_{t-1}) - z_i(\theta_{t-1})\big\|^2 \\
\leqslant \quad & \frac{2\gamma^2}{N}\sum_{i=1}^{N}\big\|f_i'(\theta_{t-1}) - f_i(\bar\theta)\big\|^2 + \frac{2\gamma^2}{N}\sum_{i=1}^{N}\big\|f_{i'}''(\bar\theta)(\theta_{t-1}-\bar\theta)\big\|^2 \\
\leqslant \quad & 2\gamma^2R^2F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + 2\gamma^2R^2[F(\bar\theta)-F(\theta_*)] \text{ from the SVRG proof },\\
& +\frac{2\gamma^2}{N}\sum_{i=1}^{N}R^2\big\|x_i^\top(\theta_{t-1}-\bar\theta)\big\|^2 \\
\leqslant \quad & 2\gamma^2R^2F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + 2\gamma^2R^2[F(\bar\theta)-F(\theta_*)] \\
& \frac{2\gamma^2R^2}{\alpha}F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + \frac{2\gamma^2R^2}{\alpha}[F(\bar\theta)-F(\theta_*)] \\
\leqslant \quad & \frac{4\gamma^2R^2}{\alpha}F'(\theta_{t-1})^\top(\theta_{t-1}-\theta_*) + \frac{4\gamma^2R^2}{\alpha}[F(\bar\theta)-F(\theta_*)]
\end{aligned}
$$

Thus, if we take the smaller step-size $\gamma = \frac{\alpha}{8R^2}$ and $T = \frac{64R^2}{\alpha\mu}$, we get the same convergence.

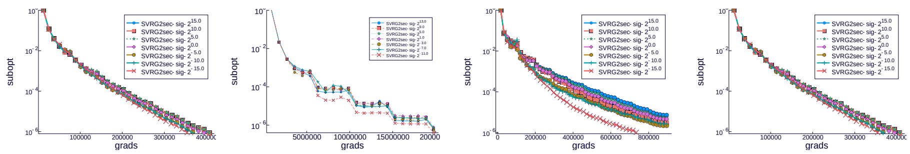
A.4 Robustness to errors in the Hessian

We assume that $z_i(\theta) = f_i'(\bar\theta) + H_i(\theta-\bar\theta)$, with a relative error $\frac{1}{N}\sum_{i=1}^{N}(f_i''(\bar\theta) - H_i)^2 \preccurlyeq R^2\eta\frac{1}{N}\sum_{i=1}^{n} f_i''(\bar\theta)$. If we take $H_i = 0$ (plain SVRG), we can take $\eta = 1$.

We assume $\frac{8\beta^2R^4}{\alpha}D^2 \leqslant L$ and $\gamma = 1/(4L)$ with $8\frac{R^2}{\alpha}\eta \leqslant L$. Then

$$
\begin{aligned}
& \frac{\gamma^2}{N}\sum_{i=1}^{N}\big\|f_i'(\theta_{t-1}) - z_i(\theta_{t-1})\big\|^2 \\
\leqslant \quad & 2\frac{\gamma^2}{N}\sum_{i=1}^{N}\big\|f_i'(\theta_{t-1}) - f_i'(\bar\theta) - f_i''(\bar\theta)(\theta_{t-1}-\bar\theta)\big\|^2 + 2\frac{\gamma^2}{N}\sum_{i=1}^{N}\big\|(H_i - f_i''(\bar\theta))(\theta_{t-1}-\bar\theta)\big\|^2 \\
\leqslant \quad & \frac{4\gamma^2\beta^2R^4}{\alpha}D^2[F(\theta_{t-1})-F(\theta_*)] + \frac{4\gamma^2\beta^2R^4}{\alpha}D^2[F(\bar\theta)-F(\theta_*)] \\
& +2\frac{\gamma^2R^2\eta}{N}\sum_{i=1}^{n}(x_i^\top(\bar\theta-\theta_{t-1}))^2 \\
\leqslant \quad & \frac{4\gamma^2\beta^2R^4}{\alpha}D^2[F(\theta_{t-1})-F(\theta_*)] + \frac{4\gamma^2\beta^2R^4}{\alpha}D^2[F(\bar\theta)-F(\theta_*)] \\
& +4\frac{\gamma^2R^2\eta}{\alpha}\bigg([F(\bar\theta)-F(\theta_*)] + [F(\theta_{t-1})-F(\theta_*)]\bigg)
\end{aligned}
$$

Thus, with the exact same proof as before (i.e., combining regular SVRG and SVRG2) we reach the desired result.



(a) `a9a` (b) `covtype` (c) `phishing` (d) `mushrooms`

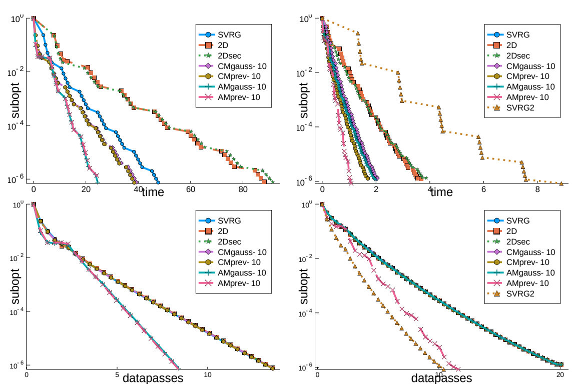
Figure 1: Performance of the SVRG2sec with different choices of σ on: (a) `a9a` (b) `covtype` (c) `phishing` (d) `mushrooms`.

B Robustness of the diagonal approximation.

Our robust secant equation has a hyperparameter, σ^2. Since the popularity of an optimization method depends as much of its ease of use as of its convergence rate, we tested the impact of σ^2 on the convergence speed. In the supplementary material we show that the impact is generally very limited and that our method is robust to the choice of σ^2. In all other experiments we set $\sigma^2 = 0.01$.

C Additional experiments

We include here the results on two additional LIBSVM datasets, *gisette* and *madelon*.



(a) `gisette` (b) `madelon`

Figure 2: Performance of various SVRG-based methods on LIBSVM test problems: (a) `gisette` (b) `madelon`.

References

[1] Sébastien Bubeck et al. “Convex optimization: Algorithms and complexity”. In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.