Technical Note Accompaniment for Babuška Forum Talk

Thomas O'Leary-Roseberry

September 27, 2024

This note serves as an accompaniment to the proof sketch in the talk. The goal of this talk is to sketch the main ideas used in the proof of $\mathcal{O}\left(\frac{1}{k}\right)$ superlinear convergence rate for Hessian-averaged Newton with adaptive gradient sampling utilizing cyclic sampling without replacement. The result in this talk is a major simplification of Theorem 3.10 in [2]. In particular, in the manuscript we are able to relax conditions of strong convexity, but this makes the details substantially more complicated. This note was written hastily, apologies for errors or typos.

Deterministic Optimization

We start with the generic deterministic unconstrained optimization problem:

$$\min_{w\in\mathbb{R}^d} f(w) \qquad \text{with minimizer} \qquad w^*. \tag{1}$$

We make the following assumptions for all $w, v \in \mathbb{R}^d$:

1. Uniform Hessian spectral bounds

$$\underbrace{\mu I \preceq \nabla^2 f(w)}_{\text{strong convexity}} \preceq LI \tag{2}$$

2. Lipschitz Hessian

$$\|\nabla^2 f(w) - \nabla^2 f(v)\| \leq M\|w-v\| \tag{3}$$

Note that 1 implies Lipschitz gradients via the mean value theorem and Cauchy–Schwarz:

$$\|\nabla f(w) - \nabla f(v)\| \leq L\|w-v\|. \tag{4}$$

Global convergence: Convergence of function values for any initial guess w_0

E.g., global linear convergence:

$$f(w_{k+1}) - f(w^*) < C\left(f(w_k) - f(w^*)\right) \tag{5}$$

with rate constant $C \in [0,1)$.

Local convergence: When iterates are sufficiently close to w^*. E.g. Q-convergence with order p.

$$\|w_{k+1} - w^*\| \leq C_p\|w_k - w^*\|^p \tag{6}$$

with rate $C_p > 0$. For Q-linear, $C_p < 1$. For strongly convex functions a global linear rate implies a local linear rate:

$$\frac{\mu}{2}\|w_k - w^*\|^2 \leq f(w_k) - f(w^*) \leq \frac{L}{2}\|w_k - w^*\|^2. \tag{7}$$

Gradient descent (GD) Global linear rate with $\alpha_k = \frac{1}{L}$, with no local improvement.

$$f(w_{k+1}) - f(w^*) \leq \left(1 - \frac{\mu}{L}\right)\left(f(w_k) - f(w^*)\right) \tag{8}$$

Newton Global linear rate with $\alpha = \frac{\mu}{L}$

$$f(w_{k+1}) - f(w^*) \leq \left(1 - \frac{\mu^2}{L^2}\right)(f(w_k) - f(w^*)). \tag{9}$$

This conservative rate is an artifact of the worst case spectral amplifications of the Hessian, in practice there is no reason to expect Newton to be worse than GD.

Stochastic and finite-sum minimization

Let ζ be a random variable, and $F(w, \zeta)$ a component function, $F_i(w) = F(w, \zeta_i)$.

$$\text{Expected risk minimization:} \quad \min_w f(w) = \mathbb{E}[F(w, \zeta)] \tag{10}$$

$$\text{Finite-Sum Minimization:} \quad \min_w f(w) = \frac{1}{n}\sum_{i=1}^{n} F_i(w) \tag{11}$$

- gradient data $X_k \subset \{1, \ldots, n\}$
- Hessian data $S_k \subset \{1, \ldots, n\}$
- $\nabla F_{X_k}(w) = \sum_{i \in X_k} \nabla F_i(w)$
- $\nabla^2 F_{S_k}(w) = \sum_{i \in S_k} \nabla^2 F_i(w)$

Hessian-averaging Inverting the subsampled Hessian leads to instabilities. At the cost of introducing a bias, we can control the variance via averaging:

$$\widehat{H}_k = \frac{1}{k}\sum_{i=1}^{k} \nabla^2 F_{S_i}(w_i). \tag{12}$$

We can use this in a fully inexact / subsampled method then as

$$\text{Subsampled Hessian-averaged Newton} \quad w_{k+1} = w_k - \alpha_k \widehat{H}_k^{-1} \nabla F_{S_k}(w_k), \tag{13}$$

where we control the gradient error via the norm condition [1], that is given a sequence $\{\theta_i\}$, we choose $|X_k|$ such that

$$\|\nabla F_{S_k}(w_k) - \nabla f(w_k)\| \leq \theta_k \|\nabla f(w_k)\|. \tag{14}$$

Superlinear convergence

- If we have a sequence e_k with a limit e^*, we say e_k converges Q-superlinearly to e^* if
$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} = 0.$$
- If there exists a sequence r_k such that $|e_k - e^*| < r_k$ and r_k converges Q-superlinearly to 0, then we say e_k converges R-superlinearly to e^*.

Main result Theorem: $\mathcal{O}\left(\frac{1}{k}\right)$ superlinear convergence of H.A. Newton (informal) Suppose w_k enters a basin of w^*, and for all S_i, $w, v \in \mathbb{R}^d$ that

- $\mu I \preceq \nabla^2 F_{S_i}(w) \preceq LI$
- $\|\nabla^2 F_{S_i}(w) - \nabla^2 F_{S_i}(v) \leq M\|w - v\|$.
- $\exists \beta_{1,H}, \beta_{2,H} < \infty$ s.t. $\|\nabla^2 F_i(w^*)\|^2 \leq \beta_{1,H}\|\nabla^2 f(w^*)\|^2 + \beta_{2,H}$.
- $\theta_k = \mathcal{O}\left(\frac{1}{k}\right)$
- Cyclic sampling w/o replacement.

Then uniformly averaged Hessian-averaged Newton w/ adaptive gradients converges superlinearly to w^* with rate $\mathcal{O}\left(\frac{1}{k}\right)$. This is a major simplification of Theorem 3.10 in [2].

Key ideas :

- We utilize cyclic sampling without replacement.
- We "mathematically move" all of the sampling error to the Hessian at the optimum: $\nabla^2 f(w^*)$
- At each epoch, the optimum sampling error is exactly zero, and the sampling errors go down at a $O\left(\frac{1}{k}\right)$ rate.
- The remaining errors are controlled through the converging iteration $w_k \to w^*$

Proof sketch

$$\begin{aligned}
\|w_{k+1} - w^*\| &= \|(w_k - w^*) - \widehat{H}_k^{-1}\nabla F_{X_k}(w_k)\| \leq \frac{1}{\mu}\|\widehat{H}_k(w_k - w^*) - \nabla F_{X_k}(w_k)\| \\
&\leq \frac{1}{\mu}\Big(\underbrace{\|\nabla^2 f(w_k)(w_k - w^*) - \nabla f(w_k)\|}_{\text{Newton error}} + \underbrace{\|(\widehat{H}_k - \nabla^2 f(w_k))(w_k - w^*)\|}_{\text{Hessian memory error}} \\
&\qquad + \underbrace{\|\nabla F_{X_k}(w_k) - \nabla f(w_k)\|}_{\text{gradient error}}\Big)
\end{aligned}$$

We will proceeed to bound each term and set up an error recursion we can bound.

$$\begin{aligned}
&\underbrace{\|\nabla^2 f(w_k)(w_k - w^*) - \nabla f(w_k)\|}_{\text{Newton error}} \\
&\quad = \left\|\nabla^2 f(w_k)(w_k - w^*) - \int_{t=0}^{1} \nabla^2 f(w_k + t(w^* - w_k))(w_k - w^*)dt\right\| \\
&\quad \leq \|w_k - w^*\| \int_{t=0}^{1} \|\nabla^2 f(w_k) - \nabla^2 f(w_k + t(w^* - w_k))\|dt \\
&\quad \leq \frac{M}{2}\|w_k - w^*\|^2
\end{aligned}$$

$$\underbrace{\|\nabla F_{X_k}(w_k) - \nabla f(w_k)\|}_{\text{gradient error}} \leq \theta_k\|\nabla f(w_k)\| = \theta_k\|\nabla f(w_k) - \nabla f(w^*)\| \leq L\theta_k\|w_k - w^*\|$$

Keep the leading order terms in mind:

$$\|w_{k+1} - w^*\| \leq \frac{1}{\mu}\Big(\underbrace{\|(\widehat{H}_k - \nabla^2 f(w_k))(w_k - w^*)\|}_{\text{Hessian memory error}} + L\theta_k\|w_k - w^*\|\frac{M}{2}\|w_k - w^*\|^2\Big)$$

The main work is then in the Hessian memory error. By judicious uses of the triangle inequality we can move all of the sampling error to w^*. See Lemma 3.5 in [2].

$$\begin{aligned}
\|(\widehat{H}_k - \nabla^2 f(w_k))(w_k - w^*)\| \leq 3M\|w_k - w^*\|^2 + \frac{M}{k}\underbrace{\left(\sum_{i=0}^{k}\|w_i - w^*\|\right)}_{\text{past iterate error}}\|w_k - w^*\| \\
\underbrace{\left\|\frac{1}{k}\sum_{i=0}^{k}\nabla^2 F_{S_i}(w^*) - \nabla^2 f(w^*)\right\|}_{\text{sampling error}}\|w_k - w^*\|
\end{aligned}$$

Proceed with past iterate error

- Before entering a local basin, we have a globally convergent iteration.
- By strong convexity we have for past iterates $\exists \rho < 1$, s.t.

$$\|w_k - w^*\|^2 \leq \frac{2}{\mu}(f(w_k) - f(w^*)) \leq \frac{2C_f}{\mu}\rho^k$$

So we can bound

$$\sum_{i=0}^{k} \|w_i - w^*\| \leq \sqrt{\frac{2C_f}{\mu}} \sum_{i=0}^{k} \sqrt{\rho}^i < C_{\text{memory}} < \infty$$

The key result that gives us the $\mathcal{O}\left(\frac{1}{k}\right)$ rate is the sampling error. See Lemma 3.8 in [2].

- Cyclic sampling without replacement.
- We assume the subsampled Hessians to have bounded error at w^*.
- We can separate each completed epoch, which sum to zero error.
- The error is only due to the remainder batches from the current epoch

$$\left\| \frac{1}{k} \sum_{i=0}^{k} \nabla^2 F_{S_i}(w^*) - \nabla^2 f(w^*) \right\| \leq \frac{C_{\text{sample}}}{k}.$$

Thus we can proceed

$$\begin{aligned}
\|(\widehat{H}_k - \nabla^2 f(w_k))(w_k - w^*)\| &\leq 3M\|w_k - w^*\|^2 + \underbrace{\frac{MC_{\text{memory}}}{k}\|w_k - w^*\|}_{\text{past iterate error}} + \underbrace{\frac{C_{\text{sampling}}}{k}}_{\text{sampling error}} \|w_k - w^*\| \\
&= 3M\|w_k - w^*\|^2 + \frac{C_{\text{avg}}}{k}\|w_k - w^*\|
\end{aligned}$$

$$\begin{aligned}
\|w_{k+1} - w^*\| &\leq \frac{1}{\mu}\left(\frac{C_{\text{avg}}}{k} + L\theta_k + \frac{7M}{2}\|w_k - w^*\|\right)\|w_k - w^*\| \\
&\leq \frac{1}{\mu}\left(\frac{C_{\text{avg}}}{k} + L\theta_k + \frac{7MC_f}{\mu}\sqrt{\rho}^k\right)\|w_k - w^*\|
\end{aligned}$$

Taking $\theta_k = \mathcal{O}\left(\frac{1}{k}\right)$, we get

$$\frac{\|w_{k+1} - w^*\|}{\|w_k - w^*\|} = \mathcal{O}\left(\frac{1}{k}\right)$$

Thus we have R-superlinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$.

- If we utilized i.i.d. sampling instead we would get the normal Monte Carlo rate $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.
- The improved rate is due to the cyclic sampling without replacement.

References

[1] M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM Journal on Scientific Computing, 34 (2012), pp. A1380–A1405.

[2] T. O'Leary-Roseberry and R. Bollapragada, *Fast Unconstrained Optimization via Hessian Averaging and Adaptive Gradient Sampling Methods*, arXiv preprint arXiv:2408.07268, (2024).