Adaptive Newton Sketch: Linear-time Optimization with Quadratic Convergence and Effective Hessian Dimensionality

Jonathan Lacotte, Yifei Wang, Mert Pilanci
lacotte@stanford.edu, wangyf18@stanford.edu, pilanci@stanford.edu
Department of Electrical Engineering, Stanford University

Introduction

Composite optimization problem

$$x^* := \operatorname*{argmin}_{x \in \mathbb{R}^d} \{f(x) := f_0(x) + g(x)\}\,.$$

(i) $f_0, g : \mathbb{R}^d \to \overline{\mathbb{R}}$ twice differentiable convex functions, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.
(ii) Forming Hessian $\nabla^2 f_0(x)$ is prohibitively expensive, while a Hessian matrix square-root $\nabla^2 f_0(x)^{1/2} \in \mathbb{R}^{n \times d}$ is available at small computational cost.
(iii) g is μ-strongly convex, i.e., $\nabla^2 g(x) \succeq \mu I_d$.

Example of the Hessian matrix square-root
$f_0(x) = \sum_{i=1}^m \ell_i(a_i^\top x)$, $\nabla^2 f_0(x)^{1/2} := \mathbf{diag}(\ell_i''(a_i^\top x)^{1/2})\,A$.
Examples of regularization functions g

- graph regularization $g(x) = \frac{1}{2}\sum_{i,j \in E}(x_i - x_j)^2$,
- ℓ_p-norms with $p > 1$,
- approximations of ℓ_1-norm.

Large-scale optimization problems of this form are very common in applications, due to the increasing dimensionality of data (e.g., genomics, medicine, high-dimensional models).
Comparison between first and second-order methods

- Newton's method enjoys **superior convergence** in both theory and practice compared to first-order methods.
- Optimal choice of first-order methods' parameters depend on unknown strong convexity and smoothness constants of problem.
- When f is *self-concordant*, then Newton's method is **invariant** to rescaling and coordinate transformations.

Newton's method. The update rule follows

$$H(x) := \nabla^2 f_0(x) + \nabla^2 g(x)\,,$$
$$x_{\text{ne}} := x - sH(x)^{-1}\nabla f(x)\,.$$

Computational issue with Newton's method: per-iteration complexity scaling as $\mathcal{O}(nd^2)$.
Newton Sketch. Our work builds on a generic method called Newton Sketch, which utilizes a random embedding of the Hessian matrix $H(x)$.
Given an embedding matrix $S \in \mathbb{R}^{m \times n}$,

$$H_S(x) := (\nabla^2 f_0(x)^{\frac{1}{2}})^\top S^\top S \nabla^2 f_0(x)^{\frac{1}{2}} + \nabla^2 g(x)\,,$$
$$x_{\text{nsk}} := x - sH_S(x)^{-1}\nabla f(x)\,.$$

Here m is a sketch size such that $m \ll n$.
For classical embeddings (e.g., sub-Gaussian, randomized orthogonal systems), a sketch size $m \asymp d$ is sufficient for the Newton sketch to achieve a linear-quadratic convergence rate with high probability (w.h.p.).

Our contribution

(i) under the assumption that g is μ-strongly convex, the scaling $m \asymp \overline{d}_\mu \log(\overline{d}_\mu)/\delta$ is sufficient for the Newton sketch to achieve a δ-accurate solution at a **quadratic** convergence rate with high probability. Here we define

$$\overline{d}_\mu := \sup_{x \in \mathcal{S}(x_0)} d_\mu(x)\,,$$

where x_0 is the initial point of our algorithm, $\mathcal{S}(x_0)$ is the sublevel set of f at x_0, and

$$d_\mu(x) := \operatorname{trace}(\nabla^2 f_0(x)(\nabla^2 f_0(x) + \mu I_d)^{-1}),$$

is the *local* effective dimension. Importantly, it always holds that $d_\mu(x) \leqslant \overline{d}_\mu \leqslant \min\{n, d\} = d$ and it can substantially smaller than the ambient dimension d.

(ii) propose an adaptive sketch size version of the effective dimension Newton sketch. Importantly, we prove that the adaptive sketch size scales in terms of $\overline{d}_\mu$. Furthermore, our adaptive method offers the possibility to the user to choose the convergence rate, from linear to quadratic.

(iii) Achieve state-of-the-art computational complexity to achieve a δ-accurate solution

$$\mathcal{O}\left(nd\log(\overline{d}_{\text{e}})\log\left(\frac{d}{\delta}\right)\log\left(\log\left(\frac{d}{\delta}\right)\right)\right).$$

Computational complexity comparisons

Algorithm	Time complexity	Sketch size	Proba.
Accelerated SVRG	$(nd + d\sqrt{\kappa n})\log(1/\delta)$	-	1
Newton method	$nd^2\log(\log(1/\delta))$	-	1
Newton sketch	$nd\log(d)\log(1/\delta)$	d	$1 - \frac{1}{d}$
Adaptive Newton sketch	$nd\log(\overline{d}_{\text{e}})\log(\frac{d}{\delta})\log(\log(\frac{d}{\delta}))$	$\frac{d}{\delta}\left(\overline{d}_{\text{e}} + \log(\frac{d}{\delta})\log(\overline{d}_{\text{e}})\right)$	$1 - \frac{1}{\overline{d}_{\text{e}}}$

Notations and background

A closed convex function $\varphi : \mathbb{R}^d \to \overline{\mathbb{R}}$ is *self-concordant* if $|\varphi'''(x)| \leqslant 2\,(\varphi''(x))^{3/2}$. This encompasses many widely used functions in practice, e.g., linear, quadratic, negative logarithm.
The choice of the sketching matrix $S \in \mathbb{R}^{m \times n}$ is critical for statistical and computational performances. Typical choices include the subsampled randomized Hadamard transform (SRHT) and the sparse Johnson-Lindenstrauss transform (SJLT).

Preliminaries

Define the Newton and approximate Newton decrements as

$$\lambda_f(x) := \left(\nabla f(x)^\top H(x)^{-1}\nabla f(x)\right)^{\frac{1}{2}},$$
$$\widetilde{\lambda}_f(x) := \left(\nabla f(x)^\top H_S(x)^{-1}\nabla f(x)\right)^{\frac{1}{2}}.$$

For a self-concordant function f, the optimality gap at any point $x \in \mathbf{dom}\, f$ is bounded in terms of the Newton decrement as

$$f(x) - f(x^*) \leqslant \lambda_f(x)^2\,.$$

Optimality gap based on approximate Newton decrements

Consider the following probability event which is critical to our convergence guarantees,

$$\boxed{\mathcal{E}_{x,m,\varepsilon} := \left\{(1 - \frac{\varepsilon}{2})I_d \preceq C_S \preceq (1 + \frac{\varepsilon}{2})I_d\right\}},$$

where $C_S := H^{-\frac{1}{2}} H_S H^{-\frac{1}{2}}$, $H \equiv H(x)$ and $H_S \equiv H_S(x)$.
Let $\varepsilon \in (0, 1/4)$ and $p \in (0, 1/2)$. It holds that $\mathbb{P}(\mathcal{E}_{x,\varepsilon,m}) \geqslant 1 - p$, provided that $m = \Omega(d_\mu(x)^2/(\varepsilon^2 p))$ for the SJLT with single nonzero element in each column, and, $m = \Omega((d_\mu(x) + \log(1/\varepsilon p)\log(d_\mu(x)/p))/\varepsilon^2)$ for the SRHT.

Closeness of Newton decrements

Let $\varepsilon \in (0, 1/4)$. Conditional on the event $\mathcal{E}_{x,m,\varepsilon}$, it holds that

$$\|v_{\text{ne}} - v_{\text{nsk}}\|_{H(x)} \leqslant \varepsilon\, \|v_{\text{ne}}\|_{H(x)}\,, \tag{1}$$
$$\sqrt{1 - \varepsilon}\,\lambda_f(x) \leqslant \widetilde{\lambda}_f(x) \leqslant \sqrt{1 + \varepsilon}\,\lambda_f(x)\,. \tag{2}$$

Adaptive Newton Sketch

We adopt the same idea as for convex quadratic objectives. Start with $m_0 = 1$, $x_0 \in \mathbb{R}^d$ and $S_0 \in \mathbb{R}^{m_0 \times n}$. At each iteration:
(i) Compute $x_{t+1} = x_t - \mu_t H_{S_t}^{-1}\nabla f(x_t)$.
(ii) Sample $S_{t+1} \in \mathbb{R}^{m_t \times n}$. Form and factorize $H_{S_{t+1}}$.
(iii) Compute improvement ratio $\widetilde{r}_t = \widetilde{\delta}_{t+1}/\widetilde{\delta}_t$ where

$$\widetilde{\delta}_t = \nabla f(x_t)^\top H_{S_t}^{-1}\nabla f(x_t)\,.$$

(iv) If $\widetilde{r}_t$ small enough, accept update x_{t+1}. Otherwise, set $x_{t+1} = x_t$, double sketch size $m_{t+1} = 2m_t$ and sample new $S_{t+1} \in \mathbb{R}^{m_{t+1} \times n}$.

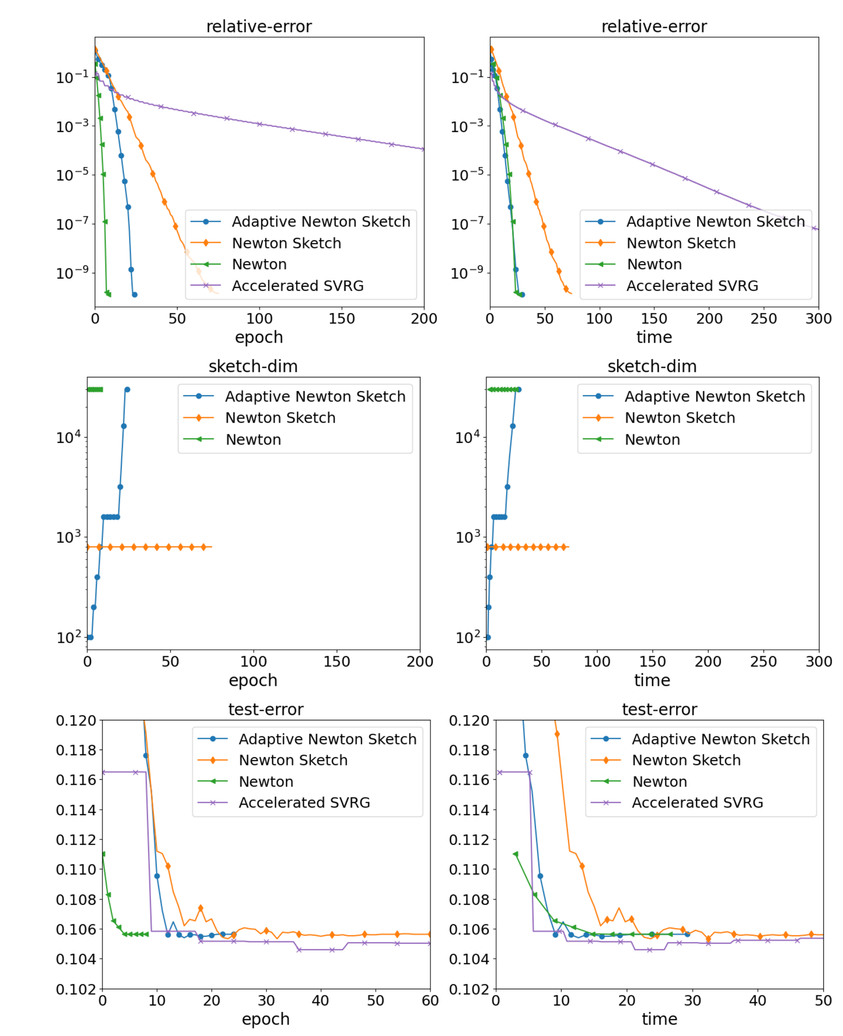
Geometric convergence guarantees of the adaptive Newton sketch

(SRHT) Let $\delta \in (0, 1/2)$. For $\tau = 1$ (quadratic rate), pick $p_0 \asymp \frac{\delta}{d}$ and assume n large enough such that $n \gtrsim \frac{d^2\overline{d}_\mu^2}{\delta^2}$. Let $\overline{m}_0$ be an initial sketch size satisfying. $\overline{m}_0 \asymp \frac{d}{\delta}\log(\frac{d}{\delta})$. Then, it holds with probability at least $1 - p_0$ that adaptive Newton sketch returns a δ-approximate solution $\widetilde{x}$ in function value (i.e., $f(\widetilde{x}) - f(x^*) \leqslant \delta$) in less than $\overline{T} = \mathcal{O}(\log(\overline{d}_\mu)\log\log(d/\delta))$ iterations, with final sketch size bounded by $2\,\overline{m} \asymp \frac{2d}{\delta}(\overline{d}_\mu + \log(\frac{d}{\delta})\log(\overline{d}_\mu))$ and with total time complexity

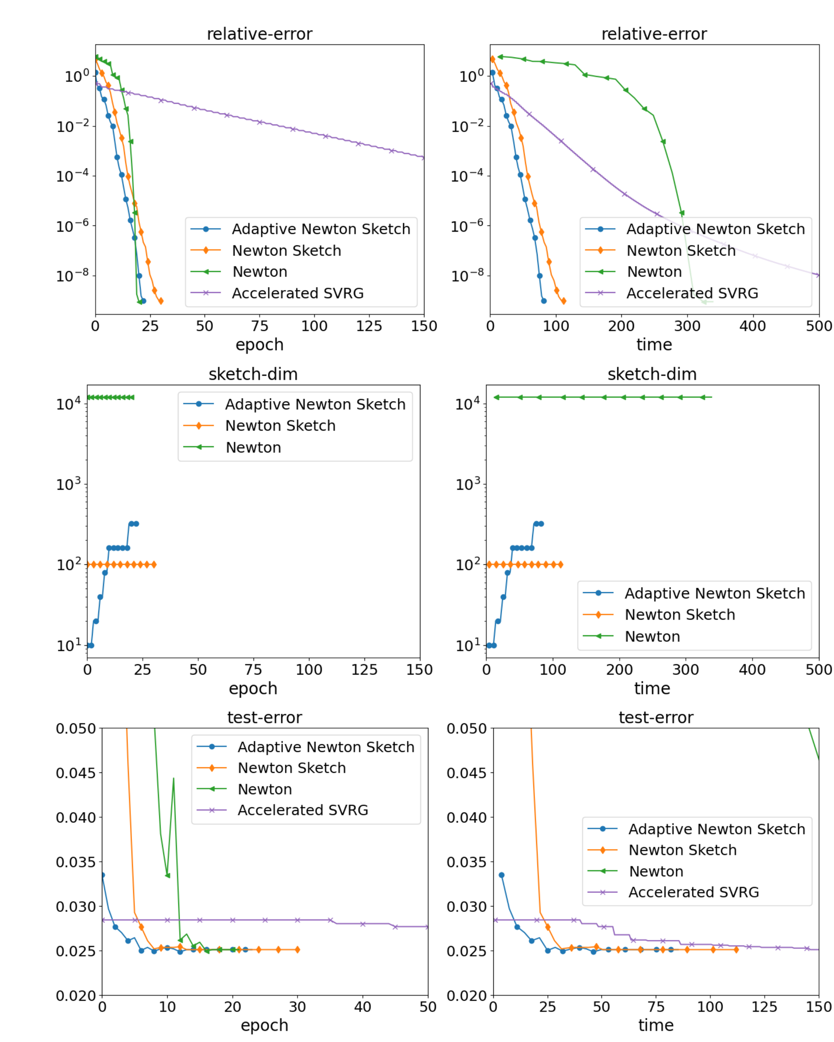
$$\overline{\mathcal{C}} = \mathcal{O}\left(nd\log(\overline{d}_\mu)\log\left(\frac{d}{\delta}\right)\log\log\left(\frac{d}{\delta}\right)\right).$$

Numerical experiments

We test on ℓ_2-regularized logistic regression problems.



MNIST. $n = 30000, d = 780, \mu = 10^{-1}$.



w7a. kernel matrix. $n = 12000, d = 12000, \mu = 10$.