Inference of Online Newton Methods with Nesterov's Accelerated Sketching

Haoxuan Wang* Xinchen Du* Sen Na

H. Milton Stewart School of Industrial and Systems Engineering (ISyE)
Georgia Institute of Technology

ICML 2026

Online inference needs uncertainty

We study stochastic optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{P}}[F(\boldsymbol{x};\xi)], \qquad \boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}).$$

Inference target

In many applications, $\boldsymbol{x}^\star$ represents the underlying true model parameter. Point estimation alone is insufficient for reliable decision-making; we need standard errors and confidence intervals.

$$\boldsymbol{w}^\top \boldsymbol{x}^\star \in \left[\boldsymbol{w}^\top \boldsymbol{x}_t \pm z_{1-q/2}\{\varphi_t \boldsymbol{w}^\top \widehat{\Sigma}_t \boldsymbol{w}\}^{0.5}\right].$$

Paper Sec. 1

Computational bottleneck for inference

Method	What it gives	Per-step cost
SGD	Efficient iterate updates, but inference procedures still require updating proper covariance matrices; delicate to ill-conditioning and noise heterogeneity of the problem.	$O(d^2)$ for inference
Exact Newton	Exploits objective curvature and can attain optimal statistical efficiency.	$O(d^3)$ solve
Sketched Newton	Approximately solves the Newton system via a sketch-and-project solver.	$O(d^2)$

Question

How does accelerated sketching affect statistical efficiency of inference in online sketched Newton methods?

Paper Sec. 1

Method in one picture

$$B_t := (1 - 1/t)\, B_{t-1} + H_{t-1}/t,$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \varphi_t\, \texttt{NASketch}(B_t, -g_t; \alpha_t, \beta_t, \gamma_t, \tau).$$

Key point

The inner randomized solver runs for a fixed number τ of sketching steps, so algorithmic randomness does not vanish asymptotically.

Paper Sec. 2

How Nesterov acceleration changes the solver

$$\begin{aligned} \boldsymbol{y}_j &= \alpha \boldsymbol{v}_j + (1-\alpha)\boldsymbol{z}_j, \\ \boldsymbol{\omega}_j &= B S_j (S_j^\top B^2 S_j)^\dagger S_j^\top (B\boldsymbol{y}_j + g), \\ \boldsymbol{z}_{j+1} &= \boldsymbol{y}_j - \boldsymbol{\omega}_j, \\ \boldsymbol{v}_{j+1} &= \beta \boldsymbol{v}_j + (1-\beta)\boldsymbol{y}_j - \gamma \boldsymbol{\omega}_j, \end{aligned}$$

$$\alpha = \frac{1}{1+\gamma\nu}, \quad \beta = 1 - \sqrt{\mu/\nu}, \quad \gamma = \frac{1}{\sqrt{\mu\nu}}.$$

Solver-level gain

Unaccelerated rate: $1-\mu$
Accelerated rate: $1-\sqrt{\mu/\nu}$

What must be tracked

Acceleration changes the inner solver into a $2d$ state–co-state recursion, inducing random parameters $(\alpha_t, \beta_t, \gamma_t)$ and a marginal sketching operator K_t.

Paper Algorithm 1 and Sec. 2.2

Last-iterate asymptotic normality

Asymptotic normality

After proving convergence of B_t, $(\alpha_t, \beta_t, \gamma_t)$, and K_t, the last Newton iterate satisfies

$$1/\sqrt{\varphi_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^\star).$$

$$[(I - K^\star) - \zeta I]\,\Sigma^\star + \Sigma^\star\,[(I - K^\star) - \zeta I] = \Gamma^\star.$$

- $K^\star$: mean marginal sketching operator.
- $\Gamma^\star$: uncertainty from random data and randomized computation.
- Exact Newton and unaccelerated sketching appear as limiting regimes.

Paper Theorem 4.3 and Proposition 4.5

Fully online covariance estimator

$$\widehat{\Sigma}_t := \frac{1}{t}\sum_{i=1}^{t}\frac{1}{\varphi_{i-1}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_t)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_t)^\top, \qquad \mathbb{E}[\|\widehat{\Sigma}_t - \Sigma^\star\|] = O(1/\sqrt{t\varphi_t}).$$

Practical inference

$$\boldsymbol{w}^\top \boldsymbol{x}^\star \in \left[\boldsymbol{w}^\top \boldsymbol{x}_t \pm z_{1-q/2}\{\varphi_t \boldsymbol{w}^\top \widehat{\Sigma}_t \boldsymbol{w}\}^{0.5}\right].$$

All summations can be updated recursively using only the accelerated Newton iterates, without requiring additional matrix-vector products. Paper Theorem 4.6

Asymptotic validity of inference

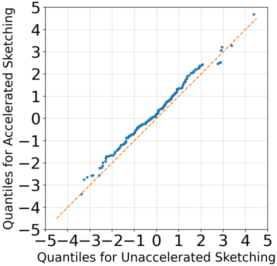
Setting, $d = 40$, Identity covariance	Exact $\tau = \infty$	$\tau = 10$	$\tau = 5$
Linear, Kaczmarz sketching	93.0%	93.0%	94.5%
Linear, Gaussian sketching	98.0%	96.0%	96.0%
Logistic, Kaczmarz sketching	94.5%	94.5%	92.5%
Logistic, Gaussian sketching	95.0%	94.0%	94.0%

Observation

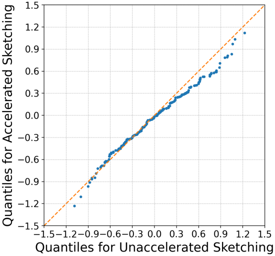
Empirical coverage rates remain close to the nominal 95% level across representative sketching settings.

Paper Table 1, 200 independent runs, 95% nominal CIs

Accelerated vs. unaccelerated covariance



Linear regression, Kaczmarz, $\tau = 5$



Logistic regression, Gaussian, $\tau = 5$

Check

Quantiles of accelerated sketched online Newton closely align with those of the unaccelerated scheme.

Paper Fig. 1 and Appendix figures

Takeaways

1. **Computation:** The method has $O(d^2)$ time and memory complexity, matching first-order methods.
2. **Inference:** The last iterate is asymptotically normal, with limiting covariance characterized by a Lyapunov equation.
3. **Practice:** A fully online covariance estimator enables practical inference using the iterates.

Thank you!