Fundamental Bias in Inverting Random Sampling Matrices with Application to Sub-sampled Newton

Chengmei Niu, Zhenyu Liao, Zenan Ling,
Huazhong University of Science and Technology, Wuhan, China
Michael W. Mahoney
ICSI, LBNL, and Department of Statistics
University of California, Berkeley, USA

ICML 2025

Outline

Classical numerical linear algebra: A computational challenge

Motivation:

- large-scale numerical linear algebra (NLA): matrix multiplication, linear systems, low-rank approximation, optimization, etc.
- high numerical precision not needed in ML; focus is on reducing computational **cost**.

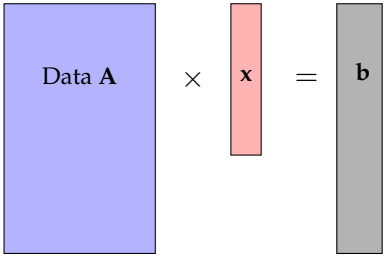
Classical numerical linear algebra: A computational challenge

Motivation:

- large-scale numerical linear algebra (NLA): matrix multiplication, linear systems, low-rank approximation, optimization, etc.
- high numerical precision not needed in ML; focus is on reducing computational **cost**.

Example of over-determined linear system: $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$

- tall data matrix with $n \gg d$, numerically challenging.

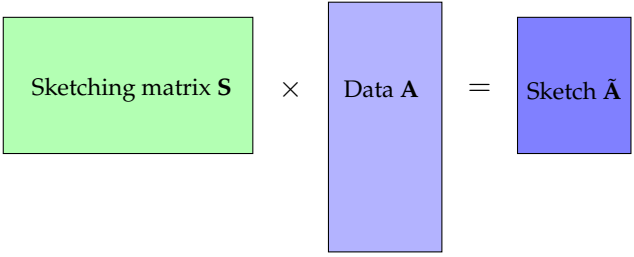


- aim to find $\mathbf{x} \in \mathbb{R}^d$ with computational complexity that does not scale rapidly with n.

Randomized methods for over-determined linear system

Idea: "sketch" the tall $\mathbf{A}$ using sketching matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ to get sketch $\tilde{\mathbf{A}} = \mathbf{SA} \in \mathbb{R}^{m \times d}$ with $m \ll n$.

- then solve the sketched system $\mathbf{SA}\tilde{\mathbf{x}} = \mathbf{Sb}$ ($m \times d$), approximate $\tilde{\mathbf{x}} \approx \mathbf{x}$.



Some commonly used random sketching techniques

- **Random sampling**: $\mathbf{S} \in \mathbb{R}^{m \times n}$ to randomly sample rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - **uniform** sampling with probability $\pi_i = 1/n$ for all $i \in \{1, \ldots, n\}$ rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$;
  - **importance** (so data-aware) sampling based on norm and leverage score (exactly or approximately) of the rows of $\mathbf{A}$ etc.

Some commonly used random sketching techniques

- **Random sampling**: $\mathbf{S} \in \mathbb{R}^{m \times n}$ to randomly sample rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - **uniform** sampling with probability $\pi_i = 1/n$ for all $i \in \{1, \ldots, n\}$ rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$;
  - **importance** (so data-aware) sampling based on norm and leverage score (exactly or approximately) of the rows of $\mathbf{A}$ etc.

Leverage score sampling [Mah11]

For $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank d with $n \geq d$, the i^{th} leverage score ℓ_i of $\mathbf{A}$, is defined as $\ell_i = \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{a}_i, i \in \{1, \ldots, n\}$. The exact leverage score sampling uses $\pi_i = \ell_i / d$.

  - In practice, use efficient approximations of the leverage scores, $\tilde{\ell}_i$, and set $\pi_i = \tilde{\ell}_i / \sum_{i=1}^{n} \tilde{\ell}_i$, which gives the approximate leverage score sampling.

Some commonly used random sketching techniques

- **Random sampling**: $\mathbf{S} \in \mathbb{R}^{m \times n}$ to randomly sample rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - **uniform** sampling with probability $\pi_i = 1/n$ for all $i \in \{1, \ldots, n\}$ rows of $\mathbf{A} \in \mathbb{R}^{n \times d}$;
  - **importance** (so data-aware) sampling based on norm and leverage score (exactly or approximately) of the rows of $\mathbf{A}$ etc.

Leverage score sampling [Mah11]

For $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank d with $n \geq d$, the i^{th} leverage score ℓ_i of $\mathbf{A}$, is defined as $\ell_i = \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{a}_i, i \in \{1, \ldots, n\}$. The exact leverage score sampling uses $\pi_i = \ell_i / d$.

  - In practice, use efficient approximations of the leverage scores, $\tilde{\ell}_i$, and set $\pi_i = \tilde{\ell}_i / \sum_{i=1}^n \tilde{\ell}_i$, which gives the approximate leverage score sampling.
- **Random projection**:
  - Gaussian and/or sub-gaussian projection
  - Sketched based on random orthonormal systems, e.g., sub-sampled randomized Hadamard/Fourier transform (SRH/FT) [AC06]
    - ⋆ for $\mathbf{SA}$ uniformly sampling the rows of $\mathbf{H}_n \mathbf{D} \mathbf{A} / \sqrt{n}$, with diagonal $\mathbf{D} \in \mathbb{R}^{n \times n}$ having i.i.d. Rademacher ± 1 entries, and $\mathbf{H}_n$ Hadamard/Fourier matrix of size n, with $\mathbf{H}_n^\top \mathbf{H}_n / n = \mathbf{I}_n$

Statistical guarantee of sketching and its inverse

For a tall data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and its sketch $\tilde{\mathbf{A}} \equiv \mathbf{S}\mathbf{A} \in \mathbb{R}^{m\times d}$, one generally needs

- **unbiased** sketch with $\mathbb{E}[\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}}] = \mathbf{A}^\top \mathbf{A}$, i.e., $\mathbb{E}[\mathbf{S}^\top \mathbf{S}] = \mathbf{I}_n$;
- $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} \simeq \mathbf{A}^\top \mathbf{A}$ in some sense with non-trivial probability for a single realization of $\mathbf{S}$.

(ε, δ)-subspace embedding [DMM06]

A sketch $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times d}$ of $\mathbf{A} \in \mathbb{R}^{n\times d}$ satisfies the subspace embedding property with error $\varepsilon \in (0,1)$ if $(1+\varepsilon)^{-1}\mathbf{A}^\top \mathbf{A} \preceq \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^\top \mathbf{A}$ holds with probability at least $1-\delta$.

Statistical guarantee of sketching and its inverse

For a tall data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and its sketch $\tilde{\mathbf{A}} \equiv \mathbf{S}\mathbf{A} \in \mathbb{R}^{m\times d}$, one generally needs

- **unbiased** sketch with $\mathbb{E}[\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}] = \mathbf{A}^\top\mathbf{A}$, i.e., $\mathbb{E}[\mathbf{S}^\top\mathbf{S}] = \mathbf{I}_n$;
- $\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} \simeq \mathbf{A}^\top\mathbf{A}$ in some sense with non-trivial probability for a single realization of $\mathbf{S}$.

(ε, δ)-subspace embedding [DMM06]

A sketch $\tilde{\mathbf{A}} \in \mathbb{R}^{m\times d}$ of $\mathbf{A} \in \mathbb{R}^{n\times d}$ satisfies the subspace embedding property with error $\varepsilon \in (0,1)$ if $(1+\varepsilon)^{-1}\mathbf{A}^\top\mathbf{A} \preceq \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^\top\mathbf{A}$ holds with probability at least $1-\delta$.

When interested in the **inverse** $(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}$,

- by (ε, δ)-subspace embedding, $(1+\varepsilon)^{-1}(\mathbf{A}^\top\mathbf{A})^{-1} \preceq (\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1} \preceq (1+\varepsilon)(\mathbf{A}^\top\mathbf{A})^{-1}$, holds with probability at least $1-\delta$;
- in general **not** unbiased $\mathbb{E}[(\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^{-1}] \not\simeq (\mathbf{A}^\top\mathbf{A})^{-1}$, due to $\mathbb{E}[1/X] \neq 1/\mathbb{E}[X]$;
- may cause large bias in practice: inaccurate LS, slow convergence in stochastic optimization, etc.

Existing work: Inversion bias under random projection

Interested in inversion bias $\mathbb{E}[(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}]$ for different random sketching $\mathbf{S} \in \mathbb{R}^{m \times n}$:

- **Gaussian S**: exactly [Haf79] $\mathbb{E}[(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}] = \frac{m}{m-d-1}(\mathbf{A}^\top \mathbf{A})^{-1}$.
- however, **beyond** Gaussian, no known exact expressions for fixed n, d, m.
- recall resolvent $\mathbf{Q}(z) = (\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A} - z\mathbf{I}_d)^{-1}$ in **RMT**, for $\mathbf{S}$ having i.i.d. entries
  - Sherman–Morrison and leave-one-out are used to approximate/compute $\mathbb{E}[\mathbf{Q}(z)]$
  - take $z = 0$ if possible

Existing work: Inversion bias under random projection

Interested in inversion bias $\mathbb{E}[(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}]$ for different random sketching $\mathbf{S} \in \mathbb{R}^{m \times n}$:

- **Gaussian S**: exactly [Haf79] $\mathbb{E}[(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}] = \frac{m}{m-d-1}(\mathbf{A}^\top \mathbf{A})^{-1}$.
- however, **beyond** Gaussian, no known exact expressions for fixed n, d, m.
- recall resolvent $\mathbf{Q}(z) = (\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A} - z\mathbf{I}_d)^{-1}$ in **RMT**, for $\mathbf{S}$ having i.i.d. entries
  - Sherman–Morrison and leave-one-out are used to approximate/compute $\mathbb{E}[\mathbf{Q}(z)]$
  - take $z = 0$ if possible

(ε, δ)-unbiased estimator

A random matrix $\tilde{\mathbf{C}}$ is an (ε, δ)-unbiased estimator of $\mathbf{C}$ if there exists an event ζ holds with probability at least $1 - \delta$ such that $(1+\varepsilon)^{-1}\mathbf{C} \preceq \mathbb{E}\left[\tilde{\mathbf{C}} \mid \zeta\right] \preceq (1+\varepsilon)\mathbf{C}$.

Existing work: Inversion bias under random projection

Interested in inversion bias $\mathbb{E}[(\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}]$ for different random sketching $\mathbf{S} \in \mathbb{R}^{m\times n}$:

- **Gaussian S**: exactly [Haf79] $\mathbb{E}[(\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}] = \frac{m}{m-d-1}(\mathbf{A}^\top\mathbf{A})^{-1}$.
- however, **beyond** Gaussian, no known exact expressions for fixed n, d, m.
- recall resolvent $\mathbf{Q}(z) = (\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A} - z\mathbf{I}_d)^{-1}$ in **RMT**, for **S** having i.i.d. entries
  - Sherman–Morrison and leave-one-out are used to approximate/compute $\mathbb{E}[\mathbf{Q}(z)]$
  - take $z = 0$ if possible

(ε, δ)-unbiased estimator

A random matrix $\tilde{\mathbf{C}}$ is an (ε, δ)-unbiased estimator of $\mathbf{C}$ if there exists an event ζ holds with probability at least $1 - \delta$ such that $(1+\varepsilon)^{-1}\mathbf{C} \preceq \mathbb{E}\left[\tilde{\mathbf{C}} \mid \zeta\right] \preceq (1+\varepsilon)\mathbf{C}$.

- Debiasing with **non-asymptotic** guarantees [Der+21a]:
  - **sub-Gaussian S**: for $m \geq Cd$, $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ with $\varepsilon = O(\sqrt{d}/m)$.
  - **LEverage Score Sparsified embedding (LESS) S** (more efficient): for $m \geq Cd\log d$, $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ with $\varepsilon = O(\sqrt{d}/m)$.

Outline

Direct row sampling

What if we (violently) randomly sample the (rows of) tall data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$?

- randomly sample m rows from n rows of $\mathbf{A}$ with replacement via probabilities $\{\pi_i\}_{i=1}^n$; form $\tilde{\mathbf{A}} = \mathbf{SA}$ and rescale each sampled row by $1/\sqrt{m\pi_i}$.
- very **sparse**: $\mathbf{S} \in \mathbb{R}^{m \times n}$ have only one nonzero entry per row

Direct row sampling

What if we (violently) randomly sample the (rows of) tall data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$?

- randomly sample m rows from n rows of $\mathbf{A}$ with replacement via probabilities $\{\pi_i\}_{i=1}^n$; form $\tilde{\mathbf{A}} = \mathbf{SA}$ and rescale each sampled row by $1/\sqrt{m\pi_i}$.
- very **sparse**: $\mathbf{S} \in \mathbb{R}^{m \times n}$ have only one nonzero entry per row

Lower bound for leverage score sampling, [Der+21a, Theorem 10]

For any $n \geq 2d \geq 4$, there exists a $\mathbf{A} \in \mathbb{R}^{n \times d}$ and approximate leverage score sampling matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ so that for any scaling factor γ, $(\gamma \mathbf{A}^\top \mathbf{S}^\top \mathbf{SA})^{-1}$ is NOT an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{A})^{-1}$ with any $\varepsilon \leq cd/m$ and $\delta \leq c(d/m)^2$, where $c > 0$ is an absolute constant.

Direct row sampling

What if we (violently) randomly sample the (rows of) tall data matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$?

- randomly sample m rows from n rows of $\mathbf{A}$ with replacement via probabilities $\{\pi_i\}_{i=1}^n$; form $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$ and rescale each sampled row by $1/\sqrt{m\pi_i}$.
- very **sparse**: $\mathbf{S} \in \mathbb{R}^{m\times n}$ have only one nonzero entry per row

Lower bound for leverage score sampling, [Der+21a, Theorem 10]

For any $n \geq 2d \geq 4$, there exists a $\mathbf{A} \in \mathbb{R}^{n\times d}$ and approximate leverage score sampling matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ so that for any scaling factor γ, $(\gamma \mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}$ is NOT an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{A})^{-1}$ with any $\varepsilon \leq cd/m$ and $\delta \leq c(d/m)^2$, where $c > 0$ is an absolute constant.

- cannot be de-biased with a **simple constant** (e.g., $\frac{m}{m-d}$ as for sub-Gaussian and LESS), at least for some approximate leverage score sampling.
- for a **win-win** for both **complexity** and **statistical property**, needs **refined** analysis to random sampling, different from projection .

Fine-grained analysis of inversion bias for random sampling

- Derive precise expression of inversion bias for general random sampling

Inversion bias for random sampling, [Niu+25, Theorem 3.1, Proposition 3.2]

Given $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank d with $n \geq d$, let $\mathbf{S} \in \mathbb{R}^{m\times n}$ be a random sampling matrix, and $\rho_{\max} \equiv \max_{1\leq i\leq n} \ell_i/(\pi_i d)$. Then, there exists $C > 0$ so that if $m \geq C\rho_{\max} d^{1+\nu}$, $\delta \leq m^{-3}$, and $\nu \geq \log_d(\log d/\delta)$,

- for diagonal matrix $\mathbf{D} = \mathrm{diag}\{D_{ii}\}_{i=1}^n$ with

$$D_{ii} = \frac{m}{m + \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{D}\mathbf{A})^{-1}\mathbf{a}_i/\pi_i}, \tag{1}$$

  $(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}$ is an (ε,δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{D}\mathbf{A})^{-1}$ for $\varepsilon = O(d^{-3\nu/2})$.

- for the de-biased sampling matrix $\check{\mathbf{S}} \in \mathbb{R}^{m\times n}$ as

$$\check{\mathbf{S}} = \mathrm{diag}\left\{\sqrt{m/(m-\ell_{i_s}/\pi_{i_s})}\right\}_{s=1}^m \cdot \mathbf{S},$$

  $(\mathbf{A}^\top \check{\mathbf{S}}^\top \check{\mathbf{S}}\mathbf{A})^{-1}$ is an (ε,δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{A})^{-1}$ for $\varepsilon = O(d^{-3\nu/2})$.

Spacial cases for scalar debiasing

Scalar debiasing under approximate leverage, [Niu+25, Corollary 3.4]

For approximate leverage sampling scheme with $\pi_i \in [\ell_i/(d\rho_{\max}), \ell_i/(d\rho_{\min})]$, where $\rho_{\min} \equiv \min_{1\le i\le n} \ell_i/(\pi_i d)$ and $\rho_{\max} \equiv \max_{1\le i\le n} \ell_i/(\pi_i d)$, there exists $C > 0$, $\nu \ge \log_d(\log d/\delta), \delta < m^{-3}$ such that for $m \ge C\rho_{\max} d^{1+\nu}$, $(\frac{m}{m-d}\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}$ is an (ε,δ)-unbiased estimator of $(\mathbf{A}^\top\mathbf{A})^{-1}$ with inversion bias $\varepsilon = O(d^{-3\nu/2} + \varepsilon_\rho d^{-\nu})$ and $\varepsilon_\rho = \max\{\rho_{\min}^{-1} - 1, 1 - \rho_{\max}^{-1}\}$.

Spacial cases for scalar debiasing

Scalar debiasing under approximate leverage, [Niu+25, Corollary 3.4]

For approximate leverage sampling scheme with $\pi_i \in [\ell_i/(d\rho_{\max}), \ell_i/(d\rho_{\min})]$, where $\rho_{\min} \equiv \min_{1\le i\le n} \ell_i/(\pi_i d)$ and $\rho_{\max} \equiv \max_{1\le i\le n} \ell_i/(\pi_i d)$, there exists $C > 0$, $\nu \ge \log_d(\log d/\delta), \delta < m^{-3}$ such that for $m \ge C\rho_{\max} d^{1+\nu}$, $(\frac{m}{m-d}\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}$ is an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{A})^{-1}$ with inversion bias $\varepsilon = O(d^{-3\nu/2} + \varepsilon_\rho d^{-\nu})$ and $\varepsilon_\rho = \max\{\rho_{\min}^{-1} - 1, 1 - \rho_{\max}^{-1}\}$.

Scalar debiasing under SRHT, [Niu+25, Corollary 3.7]

For Sub-sampled randomized Walsh–Hadamard transform (SRHT) of $\mathbf{A}$ is given by $\tilde{\mathbf{A}}_{\text{SRHT}} = \mathbf{S}\mathbf{H}_n\mathbf{D}_n\mathbf{A}/\sqrt{n} \in \mathbb{R}^{m\times n}$, then there exists $C > 0, \nu \ge 0, n\exp(-d) < \delta < m^{-3}$ such that for $m \ge C\rho_{\max} d^{1+\nu}$, $(\frac{m}{m-d}\tilde{\mathbf{A}}_{\text{SRHT}}^\top \tilde{\mathbf{A}}_{\text{SRHT}})^{-1}$ is an (ε, δ)-unbiased estimator of $(\mathbf{A}^\top \mathbf{A})^{-1}$ with inversion bias $\varepsilon = O(d^{-3\nu/2} + \rho_{\max}^{-1}\sqrt{\log(n/\delta)} d^{-\nu-1/2})$.

Outline

Application to de-biased sub-sampled Newton

- Consider optimization problem: $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathcal{C}} F(\boldsymbol{\beta})$, for some smooth function $F\colon \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{C} \subseteq \mathbb{R}^d$ a convex set
- F has Lipschitz continuous Hessian with Lipschitz constant L, where $\mathbf{H}_t(\boldsymbol{\beta}_t) \in \mathbb{R}^{d\times d}$ can be decomposed as

$$\mathbf{H}_t(\boldsymbol{\beta}_t) = \mathbf{A}(\boldsymbol{\beta}_t)^\top \mathbf{A}(\boldsymbol{\beta}_t), \tag{2}$$

  with $\mathbf{A}(\boldsymbol{\beta}_t) \in \mathbb{R}^{n\times d}$.
- Evaluate the local convergence rate of the following *de-biased* SSN iterations:

$$\check{\boldsymbol{\beta}}_{t+1} = \check{\boldsymbol{\beta}}_t - \mu_t \left(\mathbf{A}(\check{\boldsymbol{\beta}}_t)^\top \check{\mathbf{S}}_t^\top \check{\mathbf{S}}_t \mathbf{A}(\check{\boldsymbol{\beta}}_t)\right)^{-1} \mathbf{g}_t, \tag{3}$$

  with $\check{\mathbf{S}}_t = \operatorname{diag}\left\{\sqrt{m/(m-\ell_{i_s}(\check{\boldsymbol{\beta}}_t)/\pi_{i_s})}\right\}_{s=1}^{m} \cdot \mathbf{S}_t$ and i_s^{th} leverage score $\ell_{i_s}(\check{\boldsymbol{\beta}}_t)$, where $\mathbf{g}_t \equiv \nabla F(\check{\boldsymbol{\beta}}_t) \in \mathbb{R}^d$ the gradient, μ_t the step size at t.

Problem independent local convergence rate of de-biased SSN

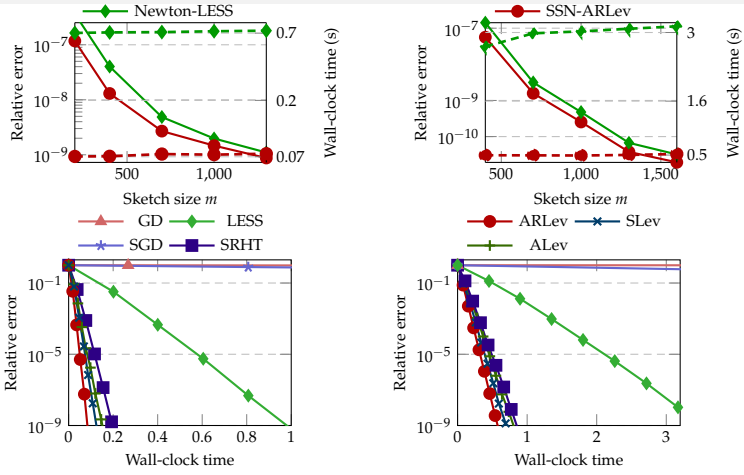
Local convergence of de-biased SSN, [Niu+25, Theorem 4.3]

For p.d. $\mathbf{A}(\boldsymbol{\beta}^*)^\top \mathbf{A}(\boldsymbol{\beta}^*) = \nabla^2 f(\boldsymbol{\beta}^*)$, there exists a neighborhood $U = \{\boldsymbol{\beta} \colon \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\mathbf{H}} < (\rho_{\max} d \sigma_{\min}/m)^{3/2}/L\}$ of $\boldsymbol{\beta}^*$ such that the de-biased SSN iteration starting from $\check{\boldsymbol{\beta}}_0 \in U$ satisfies, step size $\mu_t = 1 - \frac{\rho_{\max}}{m/d + \rho_{\max}}$, $m \geq C\rho_{\max} d^{1+\nu}$, and $\nu \geq \log_d(\log(dT/\delta))$ that

$$\left(\mathbb{E}_\zeta \left[\frac{\|\check{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*\|_{\mathbf{H}}}{\|\check{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|_{\mathbf{H}}} \right] \right)^{1/T} \leq \frac{\rho_{\max} d}{m}(1+\varepsilon), \tag{4}$$

holds for $\varepsilon = O(d^{-\nu/2})$ and conditioned on an event ζ that happens with probability at least $1-\delta$. Here, $\sigma_{\min}$ is the smallest singular value of $\mathbf{H} \equiv \mathbf{A}(\boldsymbol{\beta}^*)^\top \mathbf{A}(\boldsymbol{\beta}^*)$, $\rho_{\max}$ is the max approximate factor for $\ell_i = \max_{1 \leq t \leq T} \ell_i(\check{\boldsymbol{\beta}}_t)$ with $\ell_i(\check{\boldsymbol{\beta}}_t)$ the leverage scores of $\mathbf{A}(\check{\boldsymbol{\beta}}_t)$.

Numerical results



(a) MNIST data (b) CIFAR-10 data

- The proposed de-biased SSN-ARLev offers the **best** convergence–complexity trade-off, outperforms **Newton-LESS** and first-order baselines.

Takeaway

- **Contributions** bridging connection bewtion **RMT** and **RandNLA**:
  - propose de-biased sampling (with replacement)
    $$\check{\mathbf{S}}=\text{diag}\left\{\sqrt{m/(m-\ell_{i_s}/\pi_{i_s})}\right\}_{s=1}^{m}\cdot\mathbf{S},$$
    such that $\mathbb{E}[(\mathbf{A}^\top\check{\mathbf{S}}^\top\check{\mathbf{S}}\mathbf{A})^{-1}]\simeq(\mathbf{A}^\top\mathbf{A})^{-1}$;
  - applying de-biased $\check{\mathbf{S}}$ to sub-sampled Newton (SSN), to establish the first problem-independent local convergence rates.
- **Future work**[1]:
  - extend debiasing to block-wise sampling (with/without replacement);
  - improve SSN convergence by ensuring unbiasedness and minimizing variance.

[1]Chengmei Niu, Zhenyu Liao. "*Debiasing Distributed Subsampled Newton via A-Optimal Block Subsampling*". Manuscript in preparation, 2025.

References

- Chengmei Niu, Zhenyu Liao, Zenan Ling, and Michael W Mahoney. "Fundamental Bias in Inverting Random Sampling Matrices with Application to Sub-sampled Newton". In: *Forty-second International Conference on Machine Learning*. 2025
- Michal Derezinski, Zhenyu Liao, Edgar Dobriban, and Michael Mahoney. "Sparse sketches with small inversion bias". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. PMLR, 15–19 Aug 2021, pp. 1467–1510
- Michal Derezinski, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. "Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 2835–2847
- Michael W. Mahoney. "Randomized Algorithms for Matrices and Data". In: *Foundations and Trends® in Machine Learning* 3.2 (2011), pp. 123–224
- Nir Ailon and Bernard Chazelle. "Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform". In: *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*. 2006, pp. 557–563
- Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. "Sampling algorithms for l 2 regression and applications". In: *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. 2006, pp. 1127–1136
- Petros Drineas, Michael W Mahoney, Shan Muthukrishnan, and Tamás Sarlós. "Faster least squares approximation". In: *Numerische mathematik* 117.2 (2011), pp. 219–249
- L. R Haff. "An Identity for the Wishart Distribution with Applications". In: *Journal of Multivariate Analysis* 9.4 (Dec. 1979), pp. 531–544

Thank you!

For further discussion or questions, feel free to reach out to the authors via email:

Chengmei Niu: chengmeiniu@hust.edu.cn
Zhenyu Liao: zhenyu_liao@hust.edu.cn