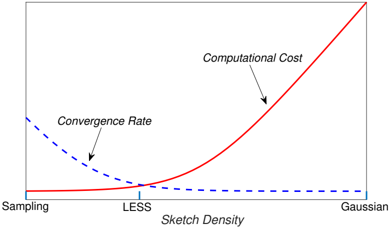
Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update

Michał Dereziński* Jonathan Lacotte† Mert Pilanci†
and Michael Mahoney‡

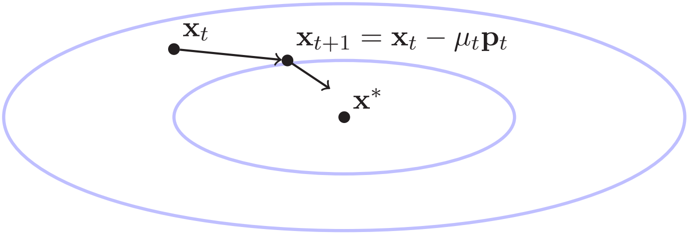
*University of Michigan, †Stanford University, ‡UC Berkeley

NeurIPS 2021



Newton's method in composite optimization

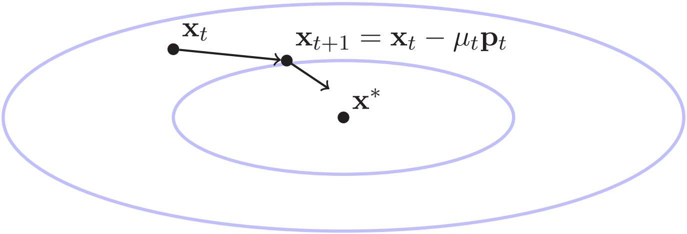
$$\text{Find:} \qquad \mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}), \qquad \text{for} \quad f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$$



Newton's method in composite optimization

Find: $\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}),$ for $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$

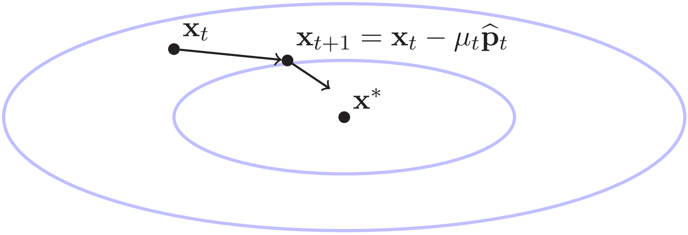
Newton step: $\mathbf{p}_t = \Big[\underbrace{\nabla^2 f(\mathbf{x}_t)}_{d \times d \text{ Hessian } \mathbf{H}} \Big]^{-1} \underbrace{\nabla f(\mathbf{x}_t)}_{d \times 1 \text{ gradient}}$



Newton's method in composite optimization

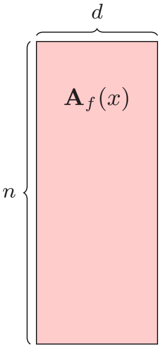
Find: $\quad \mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}), \quad$ for $\quad f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$

Newton estimate: $\quad \widehat{\mathbf{p}}_t = \Big[\underbrace{\nabla^2 \hat{f}(\mathbf{x}_t)}_{\text{Hessian estimate } \widehat{\mathbf{H}}} \Big]^{-1} \underbrace{\nabla f(\mathbf{x}_t)}_{d\times 1 \text{ gradient}}$



Computing the Hessian

$$\nabla^2 f(\mathbf{x}) = \sum_{i=1}^{n} \nabla^2 f_i(\mathbf{x}) = \overbrace{\mathbf{A}_f(\mathbf{x})^\top \mathbf{A}_f(\mathbf{x})}^{\text{Cost: } O(nd^2)}$$



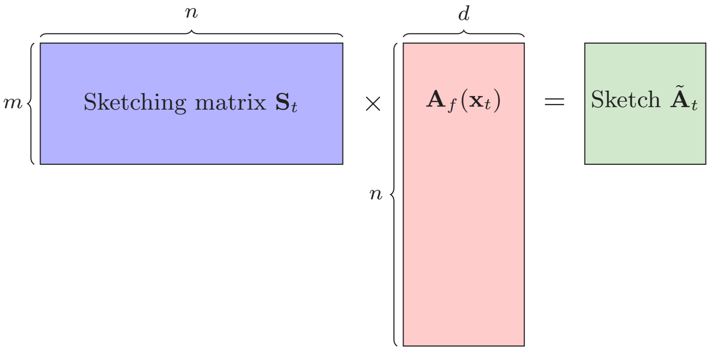
Example: Generalized Linear Model

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\phi_i^\top \mathbf{x}),$$

$$\nabla^2 f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \ell_i''(\phi_i^\top \mathbf{x}) \phi_i \phi_i^\top$$

Newton Sketch [PW17]

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{\mathbf{x}}_t - \mu_t \Big(\overbrace{\mathbf{A}_f(\widetilde{\mathbf{x}}_t)^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{A}_f(\widetilde{\mathbf{x}}_t)}^{\tilde{\mathbf{A}}_t^\top \tilde{\mathbf{A}}_t \approx \nabla^2 f(\mathbf{x}_t)} \Big)^{-1} \nabla f(\widetilde{\mathbf{x}}_t)$$



Example 1: Gaussian Newton Sketch

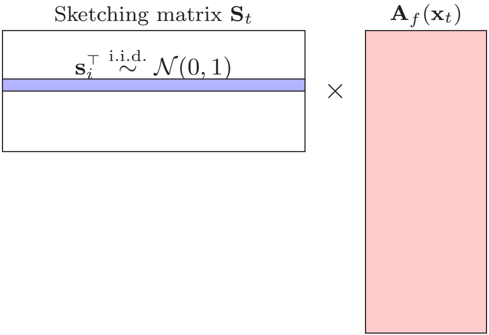
Sketching matrix $\mathbf{S}_t$ has i.i.d. Gaussian entries

Pros

- Strong convergence
- Robust to the worst case

Cons

- Computationally expensive



Extension: Sub-gaussian embeddings, e.g., with i.i.d. random sign entries

Example 2: Sub-Sampled Newton

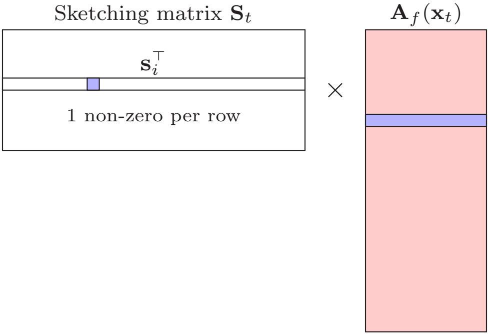
Randomly select m rows of $\mathbf{A}_f(\mathbf{x}_t)$

Pros

- Computationally cheap

Cons

- Weaker convergence
- Sensitive to the worst case

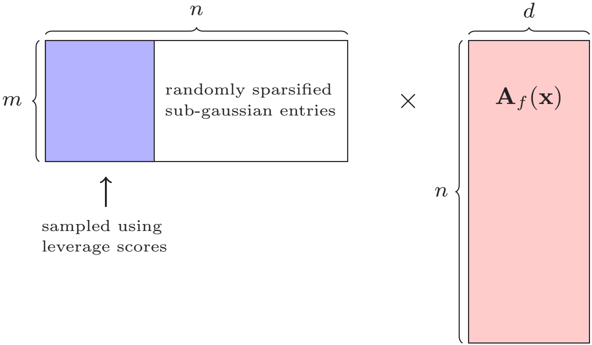


Extension: Importance sampling, e.g., according to leverage scores

LESS Embeddings: Fast Gaussian-like Sketches

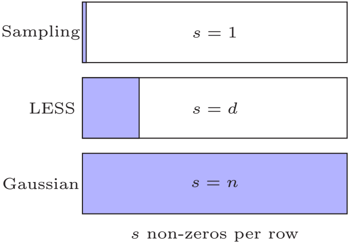
LEverage Score Sparsified (LESS) Embeddings:

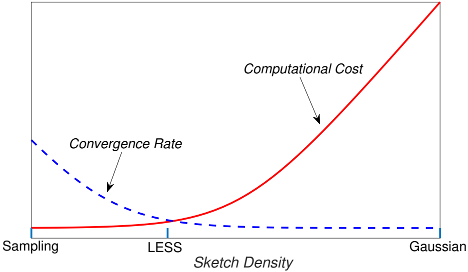
Leverage Score Sampling $+$ *Sparse Embedding Matrices*



Introduced by [DLDM21] *"Sparse sketches with small inversion bias"*, COLT'21.

Newton-LESS: Sparsity without trade-offs

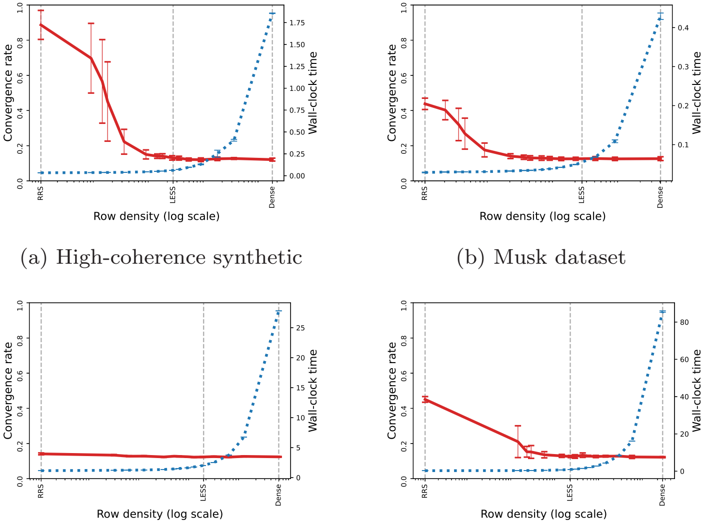




$$\text{Convergence Rate} = \left(\mathbb{E}\,\frac{\|\Delta_T\|^2}{\|\Delta_0\|^2}\right)^{1/T} \text{ where } \Delta_t = \widetilde{\mathbf{x}}_t - \mathbf{x}^*$$

$$\text{Computational Cost} = \underbrace{O(mds)}_{\text{sketch}} + \underbrace{O(md^2)}_{\text{Hessian}} + \underbrace{O(nd)}_{\text{gradient}}$$

Same plot on real data



(a) High-coherence synthetic

(b) Musk dataset

(c) CIFAR-10 dataset

(d) WESAD dataset

Main result: Problem-independent local convergence

Assumptions: Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x}^*)$ is positive definite and f is (a) self-concordant, or (b) has a Lipschitz continuous Hessian.

Sketching matrix: Gaussian, sub-Gaussian, or LESS embedding with sketch size m at least $Cd\log(dT/\delta)$

Theorem

There is a neighboorhood U containing $\mathbf{x}^*$ such that if $\widetilde{\mathbf{x}}_0 \in U$, then we can choose step size μ_t so that:

$$\left(\mathbb{E}_\delta \frac{\|\Delta_T\|_{\mathbf{H}}^2}{\|\Delta_0\|_{\mathbf{H}}^2}\right)^{1/T} \approx_\epsilon \frac{d}{m} \qquad \textit{for} \quad \epsilon = O\Big(\frac{1}{\sqrt{d}}\Big)$$

$\mathbb{E}_\delta$ is expectation conditioned on a $1-\delta$ probability event;

$\|\mathbf{v}\|_{\mathbf{H}} = \sqrt{\mathbf{v}^\top \mathbf{H}\mathbf{v}}$; $\quad a \approx_\epsilon b$ means that $(1-\epsilon)\cdot b \le a \le (1+\epsilon)\cdot b$

Main result: Problem-independent local convergence

Assumptions: Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x}^*)$ is positive definite and f is (a) self-concordant, or (b) has a Lipschitz continuous Hessian.

Sketching matrix: Gaussian, sub-Gaussian, or LESS embedding with sketch size m at least $Cd\log(dT/\delta)$

Theorem

There is a neighboorhood U containing $\mathbf{x}^*$ such that if $\widetilde{\mathbf{x}}_0 \in U$, then we can choose step size μ_t so that:

$$\left(\mathbb{E}_\delta \frac{\|\Delta_T\|_{\mathbf{H}}^2}{\|\Delta_0\|_{\mathbf{H}}^2}\right)^{1/T} \approx_\epsilon \frac{d}{m} \qquad \textit{for} \quad \epsilon = O\Big(\frac{1}{\sqrt{d}}\Big)$$

$\mathbb{E}_\delta$ is expectation conditioned on a $1-\delta$ probability event;

$\|\mathbf{v}\|_{\mathbf{H}} = \sqrt{\mathbf{v}^\top \mathbf{H}\mathbf{v}}$; $\qquad a \approx_\epsilon b$ means that $(1-\epsilon)\cdot b \le a \le (1+\epsilon)\cdot b$

Main result: Problem-independent local convergence

Assumptions: Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x}^*)$ is positive definite and f is (a) self-concordant, or (b) has a Lipschitz continuous Hessian.

Sketching matrix: Gaussian, sub-Gaussian, or LESS embedding with sketch size m at least $Cd\log(dT/\delta)$

Theorem

There is a neighboorhood U containing $\mathbf{x}^*$ such that if $\widetilde{\mathbf{x}}_0 \in U$, then we can choose step size μ_t so that:

$$\left(\mathbb{E}_\delta \frac{\|\Delta_T\|_{\mathbf{H}}^2}{\|\Delta_0\|_{\mathbf{H}}^2}\right)^{1/T} \approx_\epsilon \frac{d}{m} \qquad \textit{for} \quad \epsilon = O\Big(\frac{1}{\sqrt{d}}\Big)$$

$\mathbb{E}_\delta$ is expectation conditioned on a $1-\delta$ probability event;

$\|\mathbf{v}\|_{\mathbf{H}} = \sqrt{\mathbf{v}^\top \mathbf{H} \mathbf{v}}$; $a \approx_\epsilon b$ means that $(1-\epsilon)\cdot b \le a \le (1+\epsilon)\cdot b$

Main result: Discussion

- Same problem-independent $(\frac{d}{m})^T$ convergence rate for LESS and Gaussian (down to lower order terms)

- Simple analytic expression for the optimal step size μ_t:

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{\mathbf{x}}_t - \underbrace{\left(1 - \tfrac{d}{m}\right)}_{\mu_t} \widehat{\mathbf{p}}_t, \quad \text{when} \quad \mathbb{E}[\widehat{\mathbf{p}}_t] \approx \mathbf{p}_t.$$

- Extension to regularized objectives $f(\mathbf{x}) = f_0(\mathbf{x}) + g(\mathbf{x})$: the convergence rate becomes *dimension-independent*,

$$\left(\mathbb{E}_\delta \frac{\|\Delta_T\|_{\mathbf{H}}^2}{\|\Delta_0\|_{\mathbf{H}}^2}\right)^{1/T} \leq_\epsilon \frac{d_{\text{eff}}}{m} \quad \text{for} \quad d_{\text{eff}} = \operatorname{tr}\left(\nabla^2 f_0(\mathbf{x}^*) \nabla^2 f(\mathbf{x}^*)^{-1}\right)$$

Comparison to prior work

- Under quadratic objectives $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, the convergence rate $(\frac{d}{m})^T$ was previously shown only for:
  1. strictly Gaussian embeddings [LP19],
  2. Subsampled Randomized Hadamard Transform (SRHT) in a high-dimensional asymptotic limit [LLDP20].

- For general objectives and fast sketching methods, e.g.:
  1. Row sampling (Leverage Scores) [DMM06],
  2. Sparse sketches (CountSketch and SJLT) [CW17],
  3. Trigonometric sketches (SRHT and SRTT) [AC09],

  the best known rate is $\left(C\log(dT/\delta)\cdot\frac{d}{m}\right)^T$ [PW17].

Note: Extra constant and logarithmic factors in the bound means no analytic expressions for the optimal step size μ_t

Analysis: Two approaches

1. Subspace embedding (most prior work)
   - Standard approximation guarantee for sketching methods
   - Leads to subopotimal convergence rates: $\left(C \log(dT/\delta) \cdot \frac{d}{m}\right)^T$

$$\mathbf{A}_f(\widetilde{\mathbf{x}}_t)^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{A}_f(\widetilde{\mathbf{x}}_t) \;\approx_\eta\; \nabla^2 f(\widetilde{\mathbf{x}}_t).$$

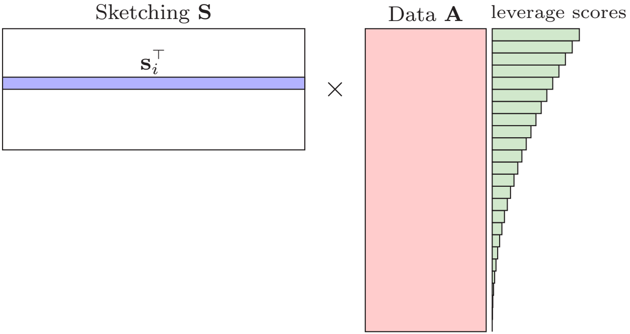
2. Method of inverse moments (this work)
   - Originally proposed for quadratic objectives [LP19]
   - Leads to precise convergence rates and optimal step sizes
   - Requires inverse moments of the sketched Hessian

$$\mathbb{E}\left[\left(\mathbf{A}_f(\widetilde{\mathbf{x}}_t)^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{A}_f(\widetilde{\mathbf{x}}_t)\right)^{-k}\right] \quad \text{for} \quad k = 1, 2$$

Comparison of sketching methods

1. Subspace embedding ✓
2. Method of inverse moments ✓

Sub-Gaussian Embedding

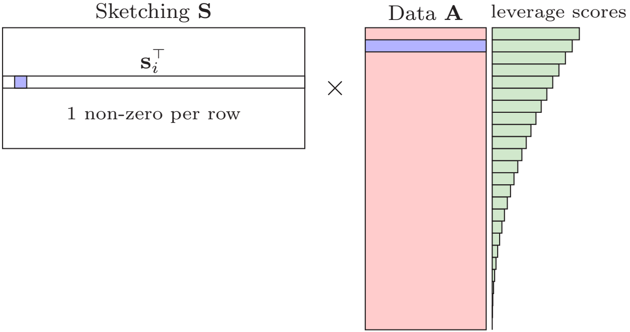


i-th leverage score: $\ell_i(\mathbf{A}) = i$-th diagonal entry of $\mathbf{A}\mathbf{A}^\dagger$

Comparison of sketching methods

1. Subspace embedding ✓
2. Method of inverse moments ✗

Leverage Score Sampling [DMM06]

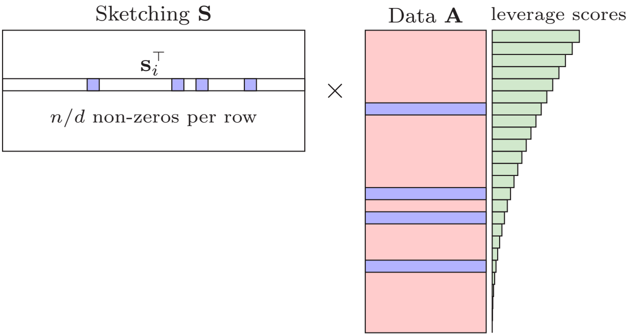


i-th leverage score: $\quad \ell_i(\mathbf{A}) \;=\; i\text{-th diagonal entry of } \mathbf{A}\mathbf{A}^\dagger$

Comparison of sketching methods

1. Subspace embedding ✓
2. Method of inverse moments ✗

Uniform Sparsification [CW13]

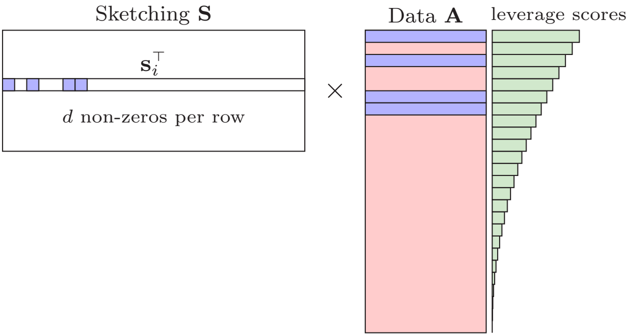


i-th leverage score: $\ell_i(\mathbf{A}) = i$-th diagonal entry of $\mathbf{A}\mathbf{A}^\dagger$

Comparison of sketching methods

1. Subspace embedding ✓
2. Method of inverse moments ✓

Leverage Score Sparsification [DLDM21]



i-th leverage score: $\quad \ell_i(\mathbf{A}) \;=\; i\text{-th diagonal entry of } \mathbf{A}\mathbf{A}^\dagger$

Implementing LESS Embeddings

1. Worst-case implementation (LESS)
   - Preprocessing cost: $O(\mathrm{nnz}(\mathbf{A}) \log n + d^3 \log d)$
     Approximating leverage scores $\ell_i(\mathbf{A})$ [DMIMW12]
   - Sketching cost: $O(md^2)$
     Sparse matrix multiplication $\mathbf{SA}$

$$\text{Cost} \quad = \quad O(\mathrm{nnz}(\mathbf{A}) \log n + md^2)$$

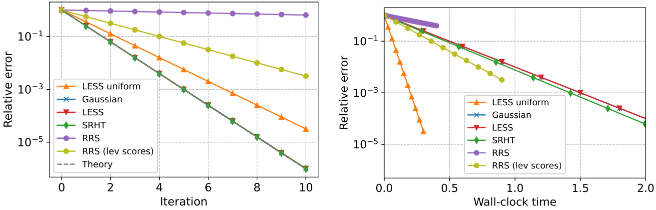
2. Practical implementation (LESS-uniform)
   - Use a uniformly sparsified sketch with αd non-zeros per row
   - If $\alpha \geq \frac{n}{d} \max_j \ell_j(\mathbf{A})$, then we recover theoretical guarantees

$$\text{Cost} \quad = \quad O(\alpha m d^2)$$

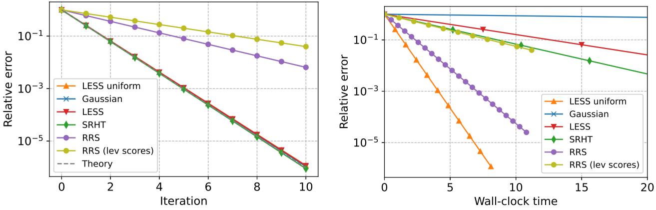
---

nnz($\mathbf{A}$) = number of non-zeros in matrix $\mathbf{A}$.

Experiments: Quadratic objective



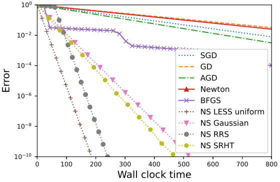
(a) High-coherence synthetic matrix



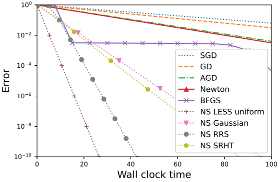
(b) WESAD dataset

We use sketch size $m = 4d$, and LESS-uniform has d non-zeros per row.

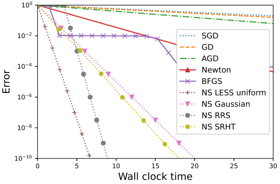
Experiments: Logistic regression



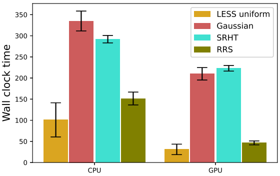
(a) WESAD dataset

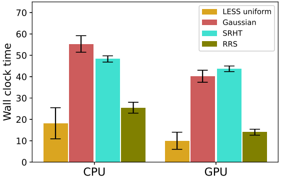


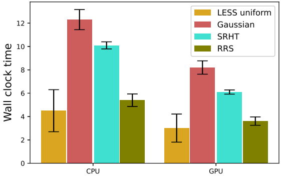
(b) CIFAR-10 dataset



(c) Musk dataset







$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-b_i \mathbf{a}_i^\top \mathbf{x})) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2 .$$

We use sketch size $m = d/2$. Bottom plots report the CPU and GPU wall-clock times to reach a 10^{-6} accurate solution.

Conclusions

- Newton-LESS: Sparsification without trade-offs
  1. Per-iteration efficiency of Sub-Sampled Newton
  2. Same convergence rate as Gaussian Newton Sketch
- Sparse sketching can beat Sub-Sampling...
  1. ...in real-world optimization tasks
  2. ...on a variety of hardware platforms
- LESS Embeddings: Fast Gaussian-like sketches
  1. Correcting the bias in distributed optimization [DLDM21]
  2. Precise convergence rates and optimal step sizes (this work)

Code available at: `https://github.com/lessketching/newtonsketch`

References I

Nir Ailon and Bernard Chazelle.
The fast Johnson–Lindenstrauss transform and approximate nearest neighbors.
SIAM Journal on computing, 39(1):302–322, 2009.

Zhidong Bai and Jack W Silverstein.
Spectral analysis of large dimensional random matrices, volume 20.
Springer, 2010.

Kenneth L. Clarkson and David P. Woodruff.
Low rank approximation and regression in input sparsity time.
In Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC '13, pages 81–90, New York, NY, USA, 2013. ACM.

Kenneth L. Clarkson and David P. Woodruff.
Low-rank approximation and regression in input sparsity time.
J. ACM, 63(6):54:1–54:45, January 2017.

Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael W Mahoney.
Sparse sketches with small inversion bias.
In Proceedings of the 34th Conference on Learning Theory, 2021.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff.
Fast approximation of matrix coherence and statistical leverage.
J. Mach. Learn. Res., 13(1):3475–3506, December 2012.

References II

Petros Drineas, Michael W Mahoney, and S Muthukrishnan.
Sampling algorithms for ℓ_2 regression and applications.
In Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, pages 1127–1136, 2006.

Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci.
Limiting spectrum of randomized Hadamard transform and optimal iterative sketching methods.
In Conference on Neural Information Processing Systems, 2020.

Jonathan Lacotte and Mert Pilanci.
Faster least squares optimization.
arXiv preprint arXiv:1911.02675, 2019.

Mert Pilanci and Martin J Wainwright.
Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence.
SIAM Journal on Optimization, 27(1):205–245, 2017.

Mark Rudelson and Roman Vershynin.
Hanson-Wright inequality and sub-gaussian concentration.
Electronic Communications in Probability, 18, 2013.