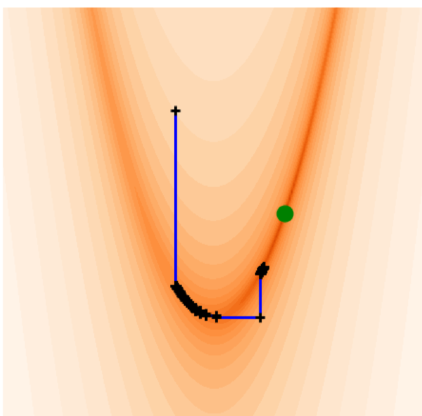


RSN: Randomized Subspace Newton

Robert M. Gower [1] Dmitry Kovalev [2] Felix Lieder [3] Peter Richtárik [2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]KAUST

[3]Heinrich-Heine-Universität Düsseldorf University

1. High Dimensional Optimization

Consider the optimization problem

$$x_* = \arg\min_{x\in\mathbb{R}^d} f(x) \ , \tag{1}$$

where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is C^2 and d is very big. This arises in training ML models with a very large number of parameters, or when data is high dimensional and acquiring data is expensive/hard.

Example: genomics, seismology, neurology and high resolution sensors in medicine.

Notation:

- Gradient & Hessian: $g(x) := \nabla f(x)$ & $\mathbf{H}(x) := \nabla^2 f(x)$
- Level set: $\mathcal{Q} := \{x \in \mathbb{R}^d : f(x) \le f(x_0)\}$
- Hessian inner product: $\langle u, v\rangle_{\mathbf{H}(x)} := \langle \mathbf{H}(x)u, v\rangle$

2. Assumptions (New)

Assumption 1: Gradient invariance:

$$g(x) \in \mathrm{Range}(\mathbf{H}(x)) \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{2}$$

Assumption 2: f is $\hat{L}$-smooth and $\hat{\mu}$-convex relative to its Hessian. That is, there exist $\hat{L} \ge \hat{\mu} > 0$ such that for all $x, y \in \mathcal{Q}$:

$$f(x) \le \underbrace{f(y) + \langle g(y), x-y\rangle + \frac{\hat{L}}{2}\|x-y\|^2_{\mathbf{H}(y)}}_{:=T(x,y)}, \tag{3}$$

$$f(x) \ge f(y) + \langle g(y), x-y\rangle + \frac{\hat{\mu}}{2}\|x-y\|^2_{\mathbf{H}(y)}. \tag{4}$$

This is a weak assumption since:

L-smoothness μ-convexity	$\Rightarrow$	c-stability [1]	$\Rightarrow$	$\hat{L}$-smoothness $\hat{\mu}$-convexity

Example: Both assumptions hold for smooth generalized linear models with L_2 regularization.

3. Newton's Method

Newton's method applied to problem (1) has the form

$$x_{k+1} = x_k - \gamma \cdot \mathbf{H}^\dagger(x_k) g(x_k) \ ,$$

where

- $\gamma > 0$ is the stepsize
- $\mathbf{H}^\dagger(x_k)$ is the Moore-Penrose pseudoinverse of $\mathbf{H}(x_k)$

Pros: Can handle curvature, invariant to coordinate transformations

Cons : Cost of each iteration is very high: $\mathcal{O}(d^3)$

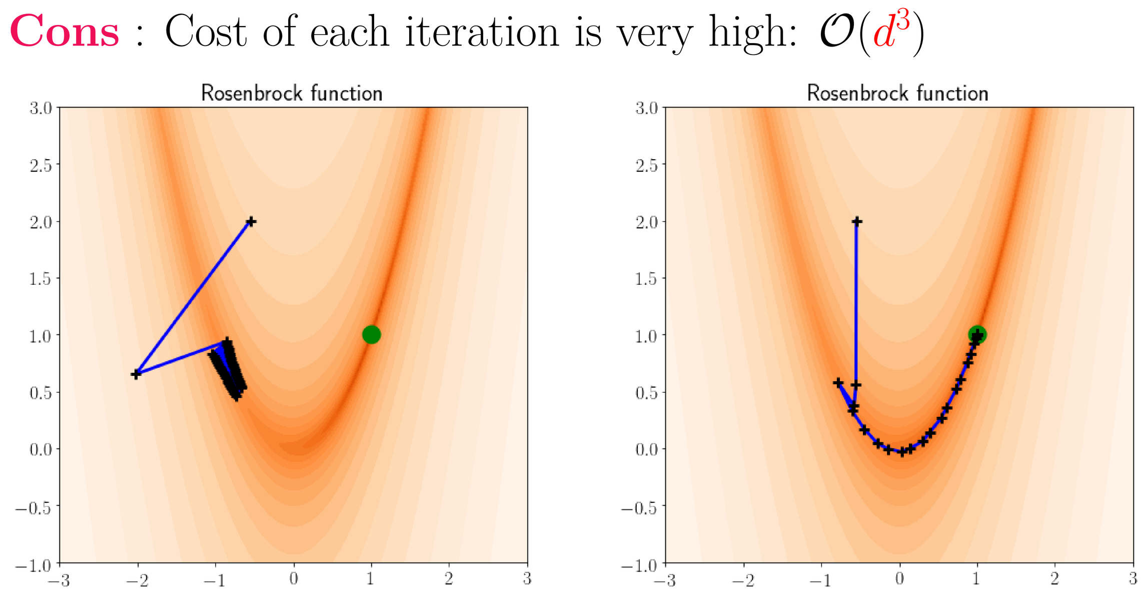
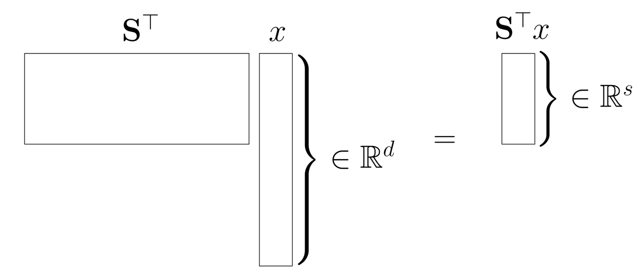


Figure: Gradient descent (left) and Newton's method (right) 50 iterations.

4. Sketching and Dimension Reduction

Let $\mathbf{S} \in \mathbb{R}^{d\times s}$ be a random matrix drawn from $\mathbf{S} \sim \mathcal{D}$.



Assumption 3: With probability 1, the sketching matrix $\mathbf{S}$ satisfies:

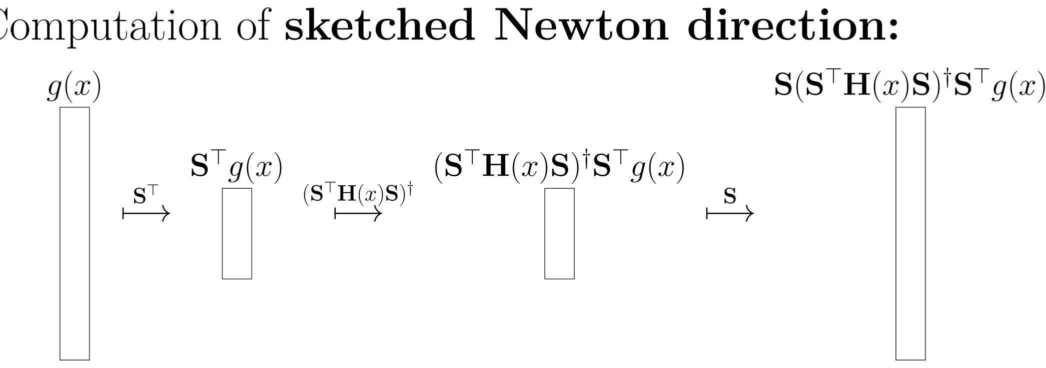
$$\mathrm{Null}\left(\mathbf{S}^\top \mathbf{H}(x)\mathbf{S}\right) = \mathrm{Null}(\mathbf{S}), \qquad \forall x \in \mathcal{Q}. \tag{5}$$

4. Randomized Subspace Newton

Algorithm 1 RSN: Randomized Subspace Newton

1. **input:** $x_0 \in \mathbb{R}^d$
2. **parameters:** $\mathcal{D}$ = distribution over random matrices
3. **for** $k = 0, 1, 2, \ldots$ **do**
4. Sample a fresh sketching matrix: $\mathbf{S}_k \sim \mathcal{D}$
5. $x_{k+1} = x_k - \frac{1}{\hat{L}}\mathbf{S}_k\left(\mathbf{S}_k^\top \mathbf{H}(x_k)\mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top g(x_k)$
6. **end for**
7. **output:** last iterate x_k

Computation of **sketched Newton direction:**



Can be computed with directional derivatives:

$$\left.\frac{\mathrm{d}f(x+\lambda\mathbf{S})}{\mathrm{d}\lambda}\right|_{\lambda=0} = \mathbf{S}^\top g(x) \qquad \left.\frac{\mathrm{d}^2 f(x+\lambda\mathbf{S})}{\mathrm{d}\lambda^2}\right|_{\lambda=0} = \mathbf{S}^\top\mathbf{H}(x)\mathbf{S}$$

Advantages of RSN:

- Uses second-order information & hence enjoys better dependence on condition number
- Enjoys global convergence theory
- Is a descent method: $f(x_{k+1}) \le f(x_k)$
- Is a feasible method: $x_k \in \mathcal{Q}$ for all $k \ge 0$
- Applicable for very large d

Example: Single Column Sketches

Let $0 \prec \mathbf{U} \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix such that $\mathbf{H}(x) \preceq \mathbf{U}$, $\forall x \in \mathbb{R}^d$. Let $\mathbf{M} = [m_1, \ldots, m_d] \in \mathbb{R}^{d\times d}$ be an invertible matrix such that $m_i^\top \mathbf{H}(x) m_i \neq 0$ for all $x \in \mathcal{Q}$ and $i = 1, \ldots, d$. If we sample according to

$$\mathrm{Prob}\left(\mathbf{S}_k = m_i\right) = p_i := \frac{m_i^\top \mathbf{U} m_i}{\mathrm{Trace}\left(\mathbf{M}^\top\mathbf{U}\mathbf{M}\right)},$$

then the update on line 5 of Algorithm 1 is given by

$$x_{k+1} = x_k - \frac{1}{\hat{L}}\frac{m_i^\top g(x_k)}{m_i^\top \mathbf{H}(x_k) m_i} m_i, \quad \text{with probability } p_i, \tag{6}$$

costs $\mathcal{O}(d)$ and has linear iteration complexity (10) given by

$$k \ge \max_{x\in\mathcal{Q}} \frac{\mathrm{Trace}\left(\mathbf{M}^\top\mathbf{U}\mathbf{M}\right)}{\lambda^+_{\min}(\mathbf{H}^{1/2}(x)\mathbf{M}\mathbf{M}^\top\mathbf{H}^{1/2}(x))}\frac{\hat{L}}{\hat{\mu}}\log\left(\frac{1}{\epsilon}\right).$$

5. RSN: Equivalent Viewpoints

1. **Minimization of $T(\cdot, x_k)$ over a random subspace:**

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d,\ \lambda\in\mathbb{R}^s} T(x, x_k) \quad \text{subject to } x = x_k + \mathbf{S}_k\lambda. \tag{7}$$

2. **Projection of the Newton direction** $n(x_k) := -\mathbf{H}^\dagger(x_k)g(x_k)$ **onto a random subspace:**

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d,\ \lambda\in\mathbb{R}^s} \left\| x - \left(x_k - \frac{1}{\hat{L}}n(x_k)\right)\right\|^2_{\mathbf{H}(x_k)} \quad \text{subject to } \quad x = x_k + \mathbf{S}_k\lambda. \tag{8}$$

3. **Projection of the current iterate** x_k **onto a sketched Newton system:**

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^d} \|x - x_k\|^2_{\mathbf{H}(x_k)} \quad \text{subject to} \quad \mathbf{S}_k^\top\mathbf{H}(x_k)(x - x_k) = -\frac{1}{\hat{L}}\mathbf{S}_k^\top g(x_k). \tag{9}$$

Remark: If $\mathrm{Range}\left(\mathbf{S}_k\right) \subset \mathrm{Range}\left(\mathbf{H}_k(x_k)\right)$, then x_{k+1} is the unique solution to (9).

6. Convergence Theory

Let $\mathbf{G}(x) := \mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{S}\left(\mathbf{S}^\top\mathbf{H}(x)\mathbf{S}\right)^\dagger\mathbf{S}\right]$ and define

$$\rho(x) := \min_{v\in\mathrm{Range}(\mathbf{H}(x))} \frac{\langle \mathbf{H}^{1/2}(x)\mathbf{G}(x)\mathbf{H}^{1/2}(x)v, v\rangle}{\|v\|_2^2}, \quad \rho := \min_{x\in\mathcal{Q}}\rho(x) \le 1.$$

Global Linear Convergence of RSN

Let $f(x_0) > f_* := \min_x f(x)$. If all assumptions hold, then

$$\mathbb{E}\left[f(x_k)\right] - f_* \le \left(1 - \rho\frac{\hat{\mu}}{\hat{L}}\right)^k (f(x_0) - f_*).$$

Consequently, given $\epsilon > 0$, if $\rho > 0$ then

$$k \ge \frac{1}{\rho}\frac{\hat{L}}{\hat{\mu}}\log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \frac{\mathbb{E}\left[f(x_k) - f_*\right]}{f(x_0) - f_*} \le \epsilon. \tag{10}$$

Sublinear Convergence of RSN

If the assumptions hold with $\hat{L} > \hat{\mu} = 0$ and

$$\mathcal{R} := \inf_{x_*\in\arg\min f}\ \sup_{x\in\mathcal{Q}} \|x - x_*\|_{\mathbf{H}(x)} < +\infty \ ,$$

and $\rho > 0$ then

$$\mathbb{E}\left[f(x_k)\right] - f_* \le \frac{2\hat{L}\mathcal{R}^2}{\rho k}. \tag{11}$$

Example: RSN includes Newton's method as a special case with $\mathbf{S}_k = \mathbf{I} \in \mathbb{R}^{d\times d}$. In this case, $\rho(x_k) \equiv 1$ and thus (10) recovers the $\hat{L}/\hat{\mu}\log(1/\epsilon)$ complexity given in [1] and (11) gives a new sublinear result.

Sufficient Condition for $\rho > 0$

If (5) holds and $\mathrm{Range}\left(\mathbf{H}(x_k)\right) \subset \mathrm{Range}\left(\mathbb{E}\left[\mathbf{S}_k\mathbf{S}_k^\top\right]\right)$, then $\rho > 0$, and $\rho = \lambda^+_{\min}\left(\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}^{1/2}(x_k)\mathbf{S}_k\left(\mathbf{S}_k^\top\mathbf{H}(x_k)\mathbf{S}_k\right)^\dagger\mathbf{S}_k^\top\mathbf{H}^{1/2}(x_k)\right]\right)$.

Example: Generalized Linear Models

Let $0 \le u \le \ell$. Let $\phi_i : \mathbb{R} \mapsto \mathbb{R}_+$ be a twice differentiable function such that

$$u \le \phi_i''(t) \le \ell, \quad \text{for } i = 1, \ldots, n. \tag{12}$$

Let $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ and $\mathbf{A} = [a_1, \ldots, a_n] \in \mathbb{R}^{d\times n}$. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is a generalized linear model when

$$f(x) = \tfrac{1}{n}\sum_{i=1}^n \phi(a_i^\top x) + \tfrac{\lambda}{2}\|x\|_2^2 \ . \tag{13}$$

f is $\hat{L}$-smooth and $\hat{\mu}$-convex relative to its Hessian with

$$\hat{L} = \frac{\ell\sigma^2_{\max}(\mathbf{A}) + n\lambda}{u\sigma^2_{\max}(\mathbf{A}) + n\lambda} \quad \text{and} \quad \hat{\mu} = \frac{u\sigma^2_{\max}(\mathbf{A}) + n\lambda}{\ell\sigma^2_{\max}(\mathbf{A}) + n\lambda}. \tag{14}$$

RSN has iteration complexity (10) given by

$$k \ge \frac{1}{\rho}\left(\frac{\ell\sigma^2_{\max}(\mathbf{A}) + n\lambda}{u\sigma^2_{\max}(\mathbf{A}) + n\lambda}\right)^2 \log\left(\frac{1}{\epsilon}\right). \tag{15}$$

7. Experiments

We compare RSN to Gradient descent (GD), accelerated gradient descent (AGD) [2] and full Newton method. For RSN we use coordinate sketches defined by $\mathbf{S}_k \in \{0,1\}^{d\times s}$, with exactly one non-zero entry per row and per column of $\mathbf{S}_k$.

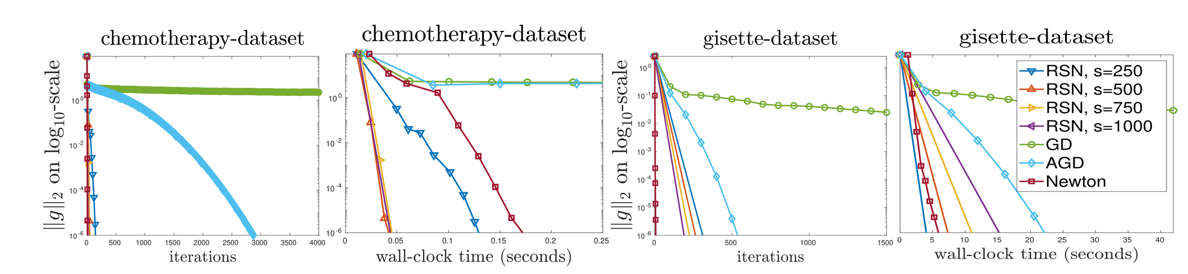


Figure: Highly dense problems, favoring RSN methods.

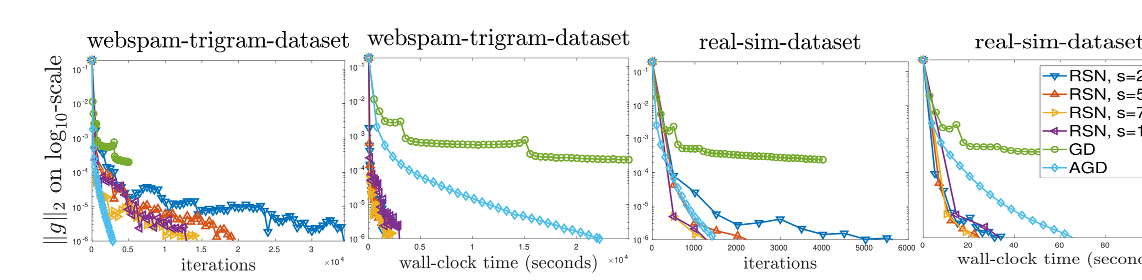


Figure: Moderately sparse problems favor the RSN method. The full Newton method is infeasible due to high dimensionality.

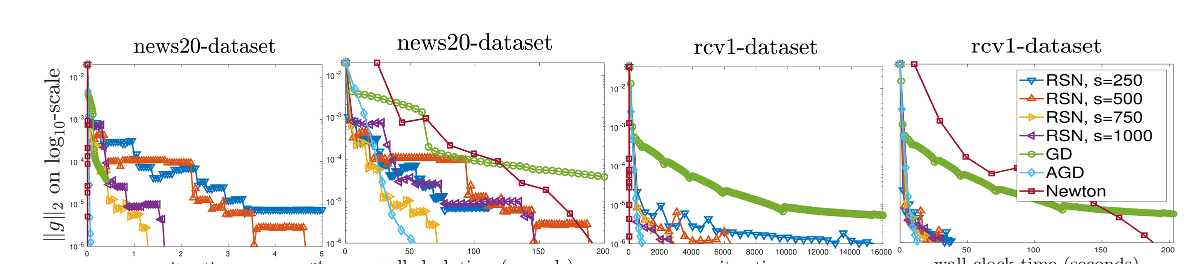


Figure: Due to extreme sparsity, accelerated gradient is competitive with the Newton type methods.

References

[1] S. P. Karimireddy, S. U. Stich, and M. Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv:1806:0041*, 2018.

[2] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer Publishing Company, Incorporated, 1 edition, 2014.