Sketch-and-Project Meets Newton Method: Global $\mathcal{O}\left(k^{-2}\right)$ Convergence with Low-Rank Updates

Anonymous Authors[1]

Abstract

In this paper, we propose the first sketch-and-project Newton method with fast $\mathcal{O}\left(k^{-2}\right)$ global convergence rate while using low-rank updates. Our method, SGN, can be viewed in three ways: i) as a sketch-and-project algorithm projecting updates of Newton method, ii) as a cubically regularized Newton method in sketched subspaces, and iii) as a damped Newton method in sketched subspaces. SGN inherits best of all three worlds: cheap iteration costs of sketch-and-project methods (up to $\mathcal{O}(1)$), state-of-the-art $\mathcal{O}\left(k^{-2}\right)$ global convergence rate of full-rank Newton-like methods and the algorithm simplicity of damped Newton methods. Finally, we demonstrate its comparable empirical performance to baseline algorithms.

Algorithm 1 SGN: Sketchy Global Newton **(new)**

Requires: Initial point $x_0 \in \mathbb{R}^d$, distribution of sketch matrices $\mathcal{D}$, constant $L_{\text{est}} \geq \sup_{\mathbf{S}\sim\mathcal{D}} L_{\mathbf{S}}$
{Choose $L_{\text{est}} \geq 1.2 \sup_{\mathbf{S}} L_{\mathbf{S}} \hat{L}_{\mathbf{S}}^2 > 0$ for global linear rate}
for $k = 0, 1, 2 \ldots$ **do**
  Sample $\mathbf{S}_k \sim \mathcal{D}$
  $\alpha_{k,\mathbf{S}} = \frac{-1+\sqrt{1+2L_{\text{est}}\left\|\nabla_{\mathbf{S}_k} f(x_k)\right\|^*_{x_k,\mathbf{S}_k}}}{L_{\text{est}}\left\|\nabla_{\mathbf{S}_k} f(x_k)\right\|^*_{x_k,\mathbf{S}_k}}$
  $x_{k+1} = x_k - \alpha_{k,\mathbf{S}} \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x_k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x_k)$
end for

1. Introduction

Second-order methods have always been a fundamental component of both scientific and industrial computing. Their origins can be traced back to the works Newton (1687), Raphson (1697), and Simpson (1740), and they have undergone extensive development throughout history (Kantorovich, 1948; Moré, 1978; Griewank, 1981). For the more historical development of classical methods, we refer the reader to (Ypma, 1995). The amount of practical applications is enormous, with over a thousand papers included in the survey Conn et al. (2000) on trust-region and quasi-Newton methods alone.

Second-order methods are highly desirable due to their invariance to rescaling and coordinate transformations, which significantly reduces the complexity of hyperparameter tuning. Moreover, this invariance allows convergence independent of the conditioning of the underlying problem. In contrast, the convergence rate of first-order methods is fundamentally dependent on the function conditioning. Moreover, first-order methods can be sensitive to variable parametrization and function scale, hence parameter tuning (e.g., step size) is often crucial for efficient execution.

On the other hand, even the simplest and most classical second-order method, Newton's method (Kantorovich, 1948), achieves an extremely fast, quadratic convergence rate (precision doubles in each iteration) (Nesterov and Nemirovski, 1994) when initialized sufficiently close to the solution. However, the convergence of the Newton method is limited only to the neighborhood of the solution. Several works, including Jarre and Toint (2016), Mascarenhas (2007), Bolte and Pauwels (2022) demonstrate that when initialized far from optimum, the line search and trust-region Newton-like method can diverge on both convex and nonconvex problems.

1.1. Demands of modern machine learning

Despite the long history of the field, research on second-order methods has been thriving to this day. Newton-like methods with a fast $\mathcal{O}\left(k^{-2}\right)$ global rate were introduced relatively recently under the name Globally Regularized Newton methods (Nesterov and Polyak, 2006; Doikov and Nesterov, 2022; Mishchenko, 2021; Hanzely et al., 2022). The main limitation of these methods is their poor scalability for modern large-scale machine learning. Large datasets with numerous features necessitate well-scalable algorithms. While tricks or inexact approximations can be used to avoid computing the inverse Hessian, simply storing the Hessian becomes impractical when the dimensionality d is large. This motivated recent developments; works Qu et al. (2016), Luo et al. (2016),

Table 1. Global convergence rates of low-rank Newton methods for convex and Lipschitz smooth functions. For simplicity, we disregard differences between various notions of smoothness. We use **fastest** full-dimensional algorithms as the baseline. For extended version, see Section 4.

Update oracle / **Update direction**	**Full-dimensional** (direction is deterministic)	**Low-rank** (direction in expectation)
Non-Newton direction	$\mathcal{O}(\mathbf{k}^{-2})$ Cubic Newton (Nesterov and Polyak, 2006) Glob. Regularized Newton (Mishchenko, 2021) (Doikov and Nesterov, 2021)	$\mathcal{O}(k^{-1})$ Stoch. Subspace Cubic Newton (Hanzely et al., 2020)
Newton direction	$\mathcal{O}(\mathbf{k}^{-2})$ Affine-Invariant Cubic Newton (Hanzely et al., 2022)	$\mathcal{O}(\mathbf{k}^{-2})$ **Sketchy Global Newton (this work)** $\mathcal{O}(k^{-1})$ Randomized Subspace Newton (Gower et al., 2019)

Gower et al. (2019), Doikov and Richtárik (2018), and Hanzely et al. (2020) propose Newton-like method operating in random low-dimensional subspaces. This approach is also known as sketch-and-project Gower and Richtárik (2015). It reduces iteration cost drastically, but for the cost of slower, $\mathcal{O}\left(k^{-1}\right)$ convergence rate (Gower et al., 2020), (Hanzely et al., 2020).

1.2. Contributions

In this work, we argue that second-order methods can be used for modern large-scale Machine Learning. We propose a **first sketch-and-project** method (Sketchy Global Newton, SGN, Algorithm 1) with fast $\mathcal{O}\left(k^{-2}\right)$ global convex convergence rate – matching global fast rate of full-dimensional Globally regularized Newton methods (as summarized in Table 1). In particular, sketching on $\mathcal{O}(1)$-dimensional subspaces leads to $\mathcal{O}\left(k^{-2}\right)$ global convex convergence with an **iteration cost** $\mathcal{O}(1)$. As a cherry on top, we additionally show **i)** local linear rate independent on the condition number, **ii)** global linear convergence under relative convexity assumption. We summarize the contributions below and in Tables 4, 2:

- **One connects all:** We present SGN through three orthogonal viewpoints: sketch-and-project method, subspace Newton method with stepsize, and Regularized Newton method. Compared to established algorithms, SGN is AICN in subspaces, SSCN in local norms, and RSN with a stepsize schedule.
- **Fast global convergence:** SGN is first low-rank method that solves **convex** functions with $\mathcal{O}\left(k^{-2}\right)$ global rate (Theorem 2). This matches state-of-the-art rates of full-rank Newton-like methods. Other sketch-and-project methods, in particular, SSCN and RSN have slower $\mathcal{O}\left(k^{-1}\right)$ rate.
- **Cheap iterations:** SGN uses τ-dimensional updates. Naively implemented, its per-iteration cost is proportional to τ^3 while full-rank Newton methods have cost proportional to d^3 and $d \gg \tau$.
- **Linear local rate:** SGN has local linear rate $\mathcal{O}\left(\frac{d}{\tau}\log\frac{1}{\varepsilon}\right)$ (Theorem 3) dependent only on the ranks of the sketching matrices. This improves over the condition-dependent linear rate of RSN or any rate of first-order methods.
- **Global linear rate:** Under $\hat{\mu}$-relative convexity, SGN achieves global linear rate $\mathcal{O}\left(\frac{L_{\text{est}}}{\rho\hat{\mu}}\log\frac{1}{\varepsilon}\right)$[1] to a neighborhood of the solution (Theorem 4).
- **Geometry and interpretability:** Update of SGN uses well-understood projections[2] of Newton method with stepsize schedule AICN. Moreover, those stochastic projections are affine-invariant and in expectation preserve direction (1). On the other hand, implicit steps of regularized Newton methods including SSCN lack geometric interpretability.
- **Algorithm simplicity:** SGN is affine-invariant and independent of the choice of the basis. This removes one parameter from potential parameter tuning. Update rule (5) is simple and explicit. Conversely, most of the fast globally-convergent Newton-like algorithms require an extra subproblem solver in each iteration.
- **Analysis:** The analysis of SGN is simple, all steps have clear geometric interpretation. On the other hand, the analysis of SSCN (Hanzely et al., 2020) is complicated as it measures distances in both l_2 norms and local norms. This not only makes it harder to understand but also leads to worse constants, which ultimately cause a slower convergence rate.

1.3. Notation

Our paper requires a nontrivial amount of notation. To facilitate reference, we will highlight new definitions in gray and theorems in light blue. We consider the optimization objective

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

where f is convex, twice differentiable, bounded from below, and potentially ill-conditioned. The number of features d is potentially large. Subspace methods use a sparse update

$$x_+ = x + \mathbf{S}h, \tag{2}$$

[1] ρ is condition number of a expected projection matrix, (17), L_{est} is constant affecting stepsize, (9).

[2] Gower et al. (2020) describes six equivalent viewpoints.

Table 2. Globally convergent Newton-like methods. For simplicity, we disregard differences and assumptions - we assume strong convexity, L-smoothness, semi-strong self-concordance and bounded level sets. We highlight the best know rates in blue.

Algorithm	Stepsize range	Affine invariant algorithm?	Iteration cost$^{(\text{dim})}$	Linear$^{(\text{lin})}$ convergence	Global convex convergence	Reference
Newton	1	✓	$\mathcal{O}\left(d^3\right)$	✗	✗	Kantorovich (1948)
Damped Newton B	$(0,1]$	✓	$\mathcal{O}\left(d^3\right)$	✗	$\mathcal{O}\left(k^{-\frac{1}{2}}\right)$	(Nesterov and Nemirovski, 1994)
AICN	$(0,1]$	✓	$\mathcal{O}\left(d^3\right)$	✗	$\mathcal{O}\left(k^{-2}\right)$	(Hanzely et al., 2022)
Cubic Newton	1	✗	$\mathcal{O}\left(d^3 \log \frac{1}{\varepsilon}\right)^{(\text{imp})}$	✗	$\mathcal{O}\left(k^{-2}\right)$	(Nesterov and Polyak, 2006)
Glob. Reg. Newton	1	✗	$\mathcal{O}\left(d^3\right)$	✗	$\mathcal{O}\left(k^{-\frac{1}{4}}\right)$	(Polyak, 2009)
Glob. Reg. Newton	1	✗	$\mathcal{O}\left(d^3\right)$	✗	$\mathcal{O}\left(k^{-2}\right)$	(Mishchenko, 2021) (Doikov and Nesterov, 2021)
Exact Newton Descent	$\frac{1}{L}^{(c)}$	✓	$\mathcal{O}\left(d^3\right)$	glob$^{(c)}$	✗	(Karimireddy et al., 2018)
RSN	$\frac{1}{L}$	✓	$\mathcal{O}\left(\tau^3\right)$	glob$^{(c)}$	$\mathcal{O}\left(k^{-1}\right)$	(Gower et al., 2019)
SSCN	1	✗	$\mathcal{O}\left(\tau^3 \log \frac{1}{\varepsilon}\right)^{(\text{imp})}$	loc	$\mathcal{O}\left(k^{-1}\right)$	(Hanzely et al., 2020)
SGN (our)	$(0,1]$	✓	$\mathcal{O}\left(\tau^3\right)$	loc + glob$^{(\text{sep})}$	$\mathcal{O}\left(k^{-2}\right)$	**This work**

(dim) d is function dimension, τ is rank of sketch matrices $\mathbf{S} \in \mathbb{R}^{d \times \tau}$. We report rate of implementation using matrix inverses.
(lin) "loc" and "glob" denotes whether algorithms have local and global linear rate (under possibly stronger assumptions).
(imp) Cubic Newton and SSCN solve implicit problem each iteration. Naively implemented, it requires $\times \log \frac{1}{\varepsilon}$ matrix inverses to approximate sufficiently in order to converge to ε-neighborhood (Hanzely et al., 2022).
(c) Authors assume c-stability or relative smoothness,implied by Lipschitz smoothness + strong convexity. (Gower et al., 2019)
(sep) Separate results for local convergence (Theorem 3) and global convergence to corresponding neighborhood (Theorem 4).

where $\mathbf{S} \in \mathbb{R}^{d \times \tau(\mathbf{S})}, \mathbf{S} \sim \mathcal{D}$ is a thin matrix and $h \in \mathbb{R}^{\tau(\mathbf{S})}$. We denote gradients and Hessians along the subspace spanned by columns of $\mathbf{S}$ as $\nabla_{\mathbf{S}} f(x) \stackrel{\text{def}}{=} \mathbf{S}^\top \nabla f(x)$ and $\nabla^2_{\mathbf{S}} f(x) \stackrel{\text{def}}{=} \mathbf{S}^\top \nabla^2 f(x) \mathbf{S}$. Also, denote any minimizer of function f as $x_* \stackrel{\text{def}}{=} \operatorname{argmin}_{x \in \mathbb{R}} f(x)$ and its value $f_* \stackrel{\text{def}}{=} f(x_*)$. We can define norms based on a symmetric positive definite matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$. For $x, g \in \mathbb{R}^d$, denote

$$\|x\|_{\mathbf{H}} \stackrel{\text{def}}{=} \langle \mathbf{H}x, x \rangle^{1/2}, \qquad \|g\|^*_{\mathbf{H}} \stackrel{\text{def}}{=} \left\langle g, \mathbf{H}^{-1} g \right\rangle^{1/2}.$$

As a special case $\mathbf{H} = \mathbf{I}$, we get l_2 norm $\|x\|_{\mathbf{I}} = \langle x, x \rangle^{1/2}$. For local Hessian norm $\mathbf{H} = \nabla^2 f(x)$, we use shorthands

$$\|h\|_x \stackrel{\text{def}}{=} \left\langle \nabla^2 f(x) h, h \right\rangle^{1/2}, \quad \|g\|^*_x \stackrel{\text{def}}{=} \left\langle g, \nabla^2 f(x)^{-1} g \right\rangle^{1/2}. \tag{3}$$

As we will be restricting iteration steps to subspaces, we will work with $\|h\|_{x,\mathbf{S}} = \|h\|_{\nabla^2_{\mathbf{S}} f(x)}$.

For a matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ and a fixed $x \in \mathbb{R}^d$, operator norm is defined by $\|\mathbf{H}\|_{op} \stackrel{\text{def}}{=} \sup_{v \in \mathbb{E}} \frac{\|\mathbf{H}v\|^*_x}{\|v\|_x}$. Note that the operator norm of Hessian in the corresponding point x is one, $\left\|\nabla^2 f(x)\right\|_{op} = 1$.

2. Three faces of the algorithm

Our algorithm combines the best of three worlds (Table 4) and we can write it in three different ways.

Theorem 1 (SGN). *If $\nabla f(x_k) \in Range(\nabla^2 f(x_k))$ then following update rules are equivalent:*

$$\bullet\, x_{k+1} = x_k + \mathbf{S}_k \operatorname*{argmin}_{h \in \mathbb{R}^d} T_{\mathbf{S}_k}(x_k, h), \tag{4}$$

$$\bullet\, x_{k+1} = x_k - \alpha_{k,\mathbf{S}_k} \mathbf{S}_k [\nabla^2_{\mathbf{S}_k} f(x_k)]^\dagger \nabla_{\mathbf{S}_k} f(x_k), \tag{5}$$

$$\bullet\, x_{k+1} = x_k - \alpha_{k,\mathbf{S}_k} \mathbf{P}^{\mathbf{S}_k}_{x_k} [\nabla^2 f(x_k)]^\dagger \nabla f(x_k), \tag{6}$$

where

$$\mathbf{P}^{\mathbf{S}}_x \stackrel{\text{def}}{=} \mathbf{S} \left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x), \tag{7}$$

$$T_{\mathbf{S}}(x, h) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), \mathbf{S}h \rangle + \tfrac{1}{2} \|\mathbf{S}h\|^2_x + \tfrac{L_{est}}{6} \|\mathbf{S}h\|^3_x, \tag{8}$$

$$\alpha_{k,\mathbf{S}} \stackrel{\text{def}}{=} \frac{-1 + \sqrt{1 + 2L_{est} \|\nabla_{\mathbf{S}} f(x_k)\|^*_{x_k,\mathbf{S}}}}{L_{est} \|\nabla_{\mathbf{S}} f(x_k)\|^*_{x_k,\mathbf{S}}}. \tag{9}$$

We call this algorithm Sketchy Global Newton, SGN (Algorithm 1), and those particular viewpoints as Regularized Newton step (4), Damped Newton step (5), and Sketch-and-project step (6).

Notice $\alpha_{k,\mathbf{S}_k} \in (0,1]$ and $\alpha_{k,\mathbf{S}_k} \xrightarrow{L_{\text{est}}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \to 0} 1$ and $\alpha_{k,\mathbf{S}_k} \xrightarrow{L_{\text{est}}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \to \infty} 0$. For SGN, we can easily transition between gradients and model differences by identity. $x_{k+1} - x_k \overset{(5)}{=} -\alpha_{k,\mathbf{S}_k}\mathbf{S}_k[\nabla^2_{\mathbf{S}_k} f(x_k)]^\dagger \nabla_{\mathbf{S}_k} f(x_k)$

2.1. Geometric properties of sketches

Matrix $\mathbf{P}^{\mathbf{S}}_x$ in (7) is a projection matrix on Range($\mathbf{S}$) w.r.t. norm $\|\cdot\|_x$ (Lemma 8) and from (6), its unbiasedness means that SGN preserves Newton's direction in expectation.

Assumption 1. *For distribution $\mathcal{D}$ there exists $\tau > 0$, s.t.*

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}^{\mathbf{S}}_x\right] = \tfrac{\tau}{d}\mathbf{I}. \tag{10}$$

Lemma 1. *Assumption 1 implies $\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}[\tau(\mathbf{S})] = \tau$.*

2.2. Invariance to affine transformations

We use assumptions invariant to the problem scale and choice of basis. An affine-invariant version of smoothness is called self-concordance, we formulate it in sketched spaces.

Definition 1. *Convex function $f \in C^3$ is $L_{\mathbf{S}}$-self-concordant in range of $\mathbf{S}$ if*

$$L_{\mathbf{S}} \overset{def}{=} \max_{x\in\mathbb{R}^d} \max_{\substack{h\in\mathbb{R}^{\tau(\mathbf{S})}\\ h\neq 0}} \frac{|\nabla^3 f(x)[\mathbf{S}h]^3|}{\|\mathbf{S}h\|_x^3}, \tag{11}$$

where $\nabla^3 f(x)[h]^3 \overset{def}{=} \nabla^3 f(x)[h,h,h]$ is 3-rd order directional derivative of f at x along $h \in \mathbb{R}^d$.

Proposition 1 (Lemma 2.2 (Hanzely et al., 2020)). *Constant $L_{\mathbf{S}}$ is determined from Range($\mathbf{S}$), and Range($\mathbf{S}$) = Range($\mathbf{S}'$) implies $L_{\mathbf{S}} = L_{\mathbf{S}'}$.*

In case $\mathbf{S} = \mathbf{I}$, Definition 1 matches definition of self-concordance and $L_{\mathbf{S}} \leq L_{\mathbf{I}}$. We will also use a slightly stronger version, semi-strong self-concordance, introduced in Hanzely et al. (2022).

Definition 2. *Convex function $f \in C^2$ is called semi-strongly self-concordant if for $\forall y, x \in \mathbb{R}^d$ holds*

$$\left\|\nabla^2 f(y) - \nabla^2 f(x)\right\|_{op} \leq L_{semi}\|y - x\|_x. \tag{12}$$

Our last convergence result is a global linear rate under relative smoothness in subspaces $\mathbf{S}$ and relative convexity. We are going to state the assumption and present rates.

Definition 3. *We call relative convexity and relative smoothness in subspace $\mathbf{S}$ positive constants $\hat{\mu}, \hat{L}_{\mathbf{S}}$ s.t. following inequalities hold $\forall x, y \in \mathcal{Q}(x_0)$ and $h \in \mathbb{R}^{\tau(\mathbf{S})}$:*

$$f(x + \mathbf{S}h) \leq f(x) + \langle \nabla_{\mathbf{S}} f(x), h\rangle + \tfrac{\hat{L}_{\mathbf{S}}}{2}\|h\|^2_{x,\mathbf{S}}, \tag{13}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\hat{\mu}}{2}\|y - x\|^2_x. \tag{14}$$

3. Convergence guarantees

We will first present the global $\mathcal{O}\left(k^{-2}\right)$ convergence rate in the convex regime. For that, denote initial level set $\mathcal{Q}(x_0) \overset{\text{def}}{=} \left\{x \in \mathbb{R}^d : f(x) \leq f(x_0)\right\}$. Lemma 4 imply that iterates of SGN stay in $\mathcal{Q}(x_0)$, $x_k \in \mathcal{Q}(x_0)\, \forall k \in \mathbb{N}$. Denote its diameter $R \overset{\text{def}}{=} \sup_{x,y\in\mathcal{Q}(x_0)} \|x - y\|_x$.

Theorem 2. *For L_{semi}-semi-strongly concordant function f with finite diameter of initial level set $\mathcal{Q}(x_0)$, $R < \infty$ and sketching matrices with Assumption 1, SGN has $\mathcal{O}\left(k^{-2}\right)$ global convergence rate,*

$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \tfrac{4d^3(f(x_0)-f_*)}{\tau^3 k^3} + \tfrac{9(\max L_{est} + L_{semi})d^2R^3}{2\tau^2k^2}. \tag{15}$$

We can state the fast local linear convergence theorem.

Theorem 3. *Let function f be $L_{\mathbf{S}}$-self-concordant in subspaces $\mathbf{S} \sim \mathcal{D}$ and expected projection matrix be unbiased (Assumption 1). For iterates of SGN $x_0, \dots, x_k$ such that $\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \leq \frac{4}{L_{\mathbf{S}_k}}$, we have local linear convergence rate*

$$\mathbb{E}\left[f(x_k) - f_*\right)] \leq \left(1 - \tfrac{\tau}{4d}\right)^k (f(x_0) - f_*) \tag{16}$$

and the local complexity of SGN is independent on the problem conditioning, $\mathcal{O}\left(\frac{d}{\tau}\log\frac{1}{\varepsilon}\right)$.

The global linear convergence rate depends on the conditioning of the scaled expected projection matrix $\mathbf{P}^{\mathbf{S}}_x$,

$$\rho(x) \overset{\text{def}}{=} [\nabla^2 f(x)]^{\frac{1}{2}}\mathbb{E}\left[\alpha_{x,\mathbf{S}}\mathbf{P}^{\mathbf{S}}_x\right][\nabla^2 f(x)]^{\dagger\frac{1}{2}}. \tag{17}$$

Theorem 4. *Let f be $L_{\mathbf{S}}$-relative smooth in subspaces $\mathbf{S}$ and $\hat{\mu}$-relative convex. Let sampling $\mathbf{S} \sim \mathcal{D}$ satisfy Null($\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}$) = Null($\mathbf{S}$) and Range($\nabla^2 f(x)$) $\subset$ Range$\left(\mathbb{E}\left[\mathbf{S}_k\mathbf{S}_k^\top\right]\right)$. Then $0 < \rho(x) \leq 1$. Denote $\rho \overset{def}{=} \min_{x\in\mathcal{Q}(x_0)}\rho(x)$ and choose parameter in stepsize $L_{est} = \sup_{\mathbf{S}\sim\mathcal{D}} \frac{9}{8}L_{\mathbf{S}}\hat{L}^2_{\mathbf{S}}$.*

While iterates $x_0, \dots, x_k$ satisfy $\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \geq \frac{4}{L_{\mathbf{S}_k}}$, then SGN has global linear rate $\mathcal{O}\left(\frac{1}{\rho\hat{\mu}}\log\frac{1}{\varepsilon}\right)$,

$$\mathbb{E}\left[f(x_k) - f_*\right] \leq \left(1 - \tfrac{4}{3}\rho\hat{\mu}\right)^k (f(x_0) - f_*). \tag{18}$$

4. Experiments

We support our theory by comparing SGN to SSCN on logistic regression empirical risk minimization on LIBSVM datasets (Chang and Lin, 2011). Figure 1 (in Appendix) shows that despite simplicity of SGN and SGN-unfavourable practical adjustments (Appendix C), SGN performs comparably to SSCN.

References

Jérôme Bolte and Edouard Pauwels. Curiosities and Counterexamples in Smooth Convex Optimization. *Mathematical Programming*, 195(1-2):553–603, 2022.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust Region Methods*. SIAM, 2000.

Nikita Doikov and Yurii Nesterov. Gradient Regularization of Newton Method with Bregman Distances. *arXiv preprint arXiv:2112.02952*, 86, 2021.

Nikita Doikov and Yurii Nesterov. Local Convergence of Tensor Methods. *Mathematical Programming*, 193: 315–336, 2022.

Nikita Doikov and Peter Richtárik. Randomized Block Cubic Newton Method. In Jennifer Dy and Andreas Krause, editors, *The 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 1290–1298, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/doikov18a.html.

Robert M Gower and Peter Richtárik. Randomized Iterative Methods for Linear Systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-Reduced Methods for Machine Learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Robert Ma Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized Subspace Newton. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 616–625. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8351-rsn-randomized-subspace-newton.pdf.

Andreas Griewank. The Modification of Newton's Method for Unconstrained Optimization by Bounding Cubic Terms. Technical report, Technical report NA/12, 1981.

Filip Hanzely, Nikita Doikov, Yurii Nesterov, and Peter Richtárik. Stochastic Subspace Cubic Newton Method. In *International Conference on Machine Learning*, pages 4027–4038. PMLR, 2020.

Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A Damped Newton Method Achieves Global $O(k^{-2})$ and Local Quadratic Convergence Rate. *Advances in Neural Information Processing Systems*, 35:25320–25334, 2022.

Florian Jarre and Philippe L Toint. Simple Examples for the Failure of Newton's Method with Line Search for Strictly Convex Minimization. *Mathematical Programming*, 158 (1):23–34, 2016.

Leonid Vital'evich Kantorovich. Functional Analysis and Applied Mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Global Linear Convergence of Newton's Method Without Strong-Convexity or Lipschitz Gradients. *arXiv:1806:0041*, 2018.

Haipeng Luo, Alekh Agarwal, Nicolo Cesa-Bianchi, and John Langford. Efficient Second Order Online Learning by Sketching. *Advances in Neural Information Processing Systems*, 29, 2016.

Walter F Mascarenhas. On the Divergence of Line Search Methods. *Computational & Applied Mathematics*, 26(1): 129–169, 2007.

Konstantin Mishchenko. Regularized Newton Method with Global $O(1/k^2)$ Convergence. *arXiv preprint arXiv:2112.02089*, 2021.

Jorge J Moré. The Levenberg-Marquardt Algorithm: Implementation and Theory. In *Numerical Analysis*, pages 105–116. Springer, 1978.

Yurii Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

Yurii Nesterov and Boris T Polyak. Cubic Regularization of Newton Method and Its Global Performance. *Mathematical Programming*, 108(1):177–205, 2006.

Isaac Newton. *Philosophiae Naturalis Principia Mathematica*. Jussu Societatis Regiae ac Typis Josephi Streater, 1687.

Roman A Polyak. Regularized Newton Method for Unconstrained Convex Optimization. *Mathematical Programming*, 120(1):125–145, 2009.

Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq. SDNA: Stochastic dual Newton ascent for empirical risk minimization. In *The 33rd International Conference on Machine Learning (ICML)*, pages 1823–1832, 2016.

Joseph Raphson. *Analysis Aequationum Universalis Seu Ad Aequationes Algebraicas Resolvendas Methodus Generalis & Expedita, Ex Nova Infinitarum Serierum Methodo, Deducta Ac Demonstrata*. Th. Braddyll, 1697.

Thomas Simpson. *Essays on Several Curious and Useful Subjects, in Speculative and Mix'd Mathematicks. Illustrated by a Variety of Examples.* Printed by H. Woodfall, jun. for J. Nourse, at the Lamb without Temple-Bar, 1740.

Tjalling J Ypma. Historical Development of the Newton–Raphson Method. *SIAM Review*, 37(4):531–551, 1995.

A. Table of notation

$A^\dagger$	Moreau pseudoinverse of A
d	dimension of problem
$f : \mathbb{R}^d \to \mathbb{R}$	optimization function
$x, x_+, x_k \in \mathbb{R}^d$	iterates
$y \in \mathbb{R}^d$	virtual iterate (for analysis only)
$h, h' \in \mathbb{R}^d$	difference between consecutive iterates
x_*	optimal model
f_*	optimal function value
$\mathcal{Q}(x_0)$	set of models with functional value less than x_0
R	diameter of $\mathcal{Q}(x_0)$
$\\|\cdot\\|_{op}$	operator norm
$\nabla_{\mathbf{S}} f, \nabla^2_{\mathbf{S}} f, \\|h\\|_{x,\mathbf{S}}$	gradient, Hessian, local norm in range $\mathbf{S}$, resp.
$\\|\cdot\\|_x$	local norm at x
$\\|\cdot\\|_x^*$	local dual norm at x
$\alpha_{k,\mathbf{S}_k}$	SGN stepsize
$T_{\mathbf{S}}(\cdot, x)$	upperbound on f based on gradient and Hessian in x
$\mathbf{S} \in \mathbb{R}^{d \times \tau(\mathbf{S})}$	randomized sketching matrix
$\tau(\mathbf{S})$	dimension of randomized sketching matrix
τ	fixed dimension constraint on $\mathbf{S}$
$L_{\mathbf{S}}$	self-concordance constant in range of $\mathbf{S}$
$\mathbf{P}_x^{\mathbf{S}}$	projection matrix on subspace $\mathbf{S}$ w.r.t. local norm at x
$\rho(x)$	condition numbers of expected projection matrix $\mathbb{E}\left[\mathbf{P}_x^{\mathbf{S}}\right]$
ρ	lower bound on condition numbers $\rho(x)$
$L_{\text{sc}}, L_{\text{semi}}$	self-concordance and semi-strong self-concordance constants, resp.
L_{est}	smoothness estimate, affects stepsize of SGN
$\hat{L}, \hat{\mu}$	relative smoothness and relative convexity constants

Table 3. Table of notation

B. Insights

B.1. Sketch matrices $\mathbf{P}_x^{\mathbf{S}}$

Note that restriction on the sketch matrices $\mathbf{P}_x^{\mathbf{S}}$ (Assumption 1) is formulated in the local norm, so it might seem restrictive. Next lemma demonstrates that such sketching matrices can be obtained from sketches with l_2-unbiased projection (which were used in (Hanzely et al., 2020)).

Lemma 2 (Construction of sketch matrix $\mathbf{S}$). *If we have a sketch matrix distribution $\tilde{\mathcal{D}}$ so that a projection on Range$(\mathbf{M}), \mathbf{M} \sim \mathcal{D}$ is unbiased in l_2 norms,*

$$\mathbb{E}_{\mathbf{M}\sim\tilde{\mathcal{D}}}\left[\mathbf{M}^\top \left(\mathbf{M}^\top \mathbf{M}\right)^\dagger \mathbf{M}\right] = \tfrac{\tau}{d}\mathbf{I}, \tag{19}$$

then distribution $\mathcal{D}$ of $\mathbf{S}$ defined as $\mathbf{S}^\top \overset{def}{=} \mathbf{M}\left[\nabla^2 f(x)\right]^{-1/2}$ (for $\mathbf{M} \sim \tilde{\mathcal{D}}$) satisfy Assumption 1,

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_x^{\mathbf{S}}\right] = \tfrac{\tau}{d}\mathbf{I}. \tag{20}$$

Matrices $\mathbf{P}_x^{\mathbf{S}}$ have easily clear contractive properties, as tated in the next lemma.

Table 4. Three approaches for second-order global minimization. We denote $x_k \in \mathbb{R}^d$ model iterates, $\mathbf{S}_k \sim \mathcal{D}$ distribution of sketch matrices with rank $\tau \ll d$, $\alpha_k, \alpha_{k,\mathbf{S}_k}$ stepsizes, $L_2, L_{\mathbf{S}}$ smoothness constants, c_{stab} Hessian stability constant. For simplicity, we disregard differences in assumptions. We report algorithm complexities when matrix inverses are naively implemented.

Orthogonal lines of work	**Sketch-and-Project** (Gower and Richtárik, 2015) (various update rules)	**Damped Newton** (Nesterov and Nemirovski, 1994) (Karimireddy et al., 2018)	**Globally Regularized Newton** [(r)] (Nesterov and Polyak, 2006) (Polyak, 2009) (Mishchenko, 2021) (Doikov and Nesterov, 2021)
Update $x_{k+1} - x_k =$	$\alpha_{k,\mathbf{S}_k} \mathbf{P}_{x_k}^{\mathbf{S}_k}$ (update(x_k)), for $\mathbf{S}_k \sim \mathcal{D}$	$\alpha_k [\nabla^2 f(x_k)]^\dagger \nabla f(x_k)$	$\operatorname{argmin}_{h \in \mathbb{R}^d} T(x_k, h)$, for $T(x,h) \stackrel{\text{def}}{=} \langle \nabla f(x), h \rangle + \frac{1}{2}\|h\|_x^2 + \frac{L_2}{6}\|h\|_2^3$
Characteristics	+ cheap, low-rank updates + global linear convergence (conditioning-dependent) - optimal rate: linear	+ affine-invariant geometry - iteration cost $\mathcal{O}(d^3)$ Fixed $\alpha_k = c_{\text{stab}}^{-1}$: + global linear convergence Schedule $\alpha_k \nearrow 1$: + local quadratic rate	+ global convex rate $\mathcal{O}(k^{-2})$ + local quadratic rate - implicit updates - iteration cost $\mathcal{O}(d^3 \log \frac{1}{\varepsilon})$
Combinations + retained benefits	**Sketch-and-Project**	**Damped Newton**	**Globally Regularized Newton**
RSN (Gower et al., 2019) Algorithm 3	✓ + iteration cost $\mathcal{O}(\tau^3)$	✓ + global rate $\mathcal{O}\left(\frac{1}{\rho}\frac{\hat{L}}{\hat{\mu}} \log \frac{1}{\varepsilon}\right)$	✗
SSCN (Hanzely et al., 2020) Algorithm 4	✓ + iteration cost $\mathcal{O}(\tau^3 \log \frac{1}{\varepsilon})$ + local rate $\mathcal{O}(\frac{d}{\tau} \log \frac{1}{\varepsilon})$	✗	✓ + global convex rate $\mathcal{O}(k^{-2})$
AICN (Hanzely et al., 2022) Algorithm 5	✗	✓ + affine-invariant geometry - no global linear rate proof [(lin)]	✓ + global convex rate $\mathcal{O}(k^{-2})$ + local quadratic rate + iteration cost $\mathcal{O}(d^3)$ + simple, explicit updates
SGN (this work) Algorithm 1	✓ + iteration cost $\mathcal{O}(\tau^3 \log \frac{1}{\varepsilon})$ + local rate $\mathcal{O}(\frac{d}{\tau} \log \frac{1}{\varepsilon})$ - quadratic rate unachievable	✓ + affine-invariant geometry + global rate $\mathcal{O}\left(\frac{1}{\rho}\frac{\hat{L}}{\hat{\mu}} \log \frac{1}{\varepsilon}\right)$	✓ + global convex rate $\mathcal{O}(k^{-2})$ + simple, explicit updates
Three descriptions of SGN	**Sketch-and-Project of Damped Newton method**	**Damped Newton in sketched subspaces**	**Affine-Invariant Cubic Newton in sketched subspaces**
Update $x_{k+1} - x_k =$	$\alpha_{k,\mathbf{S}_k} \mathbf{P}_{x_k}^{\mathbf{S}_k} [\nabla^2 f(x_k)]^\dagger \nabla f(x_k)$	$\alpha_{k,\mathbf{S}_k} \mathbf{S}_k [\nabla_{\mathbf{S}_k} f(x_k)]^\dagger \nabla_{\mathbf{S}_k} f(x_k)$	$\mathbf{S}_k \operatorname{argmin}_{h \in \mathbb{R}^d} T_{\mathbf{S}_k}(x_k, h)$, for $T_{\mathbf{S}}(x,h) \stackrel{\text{def}}{=} \langle \nabla f(x), \mathbf{S}h \rangle + \frac{1}{2}\|\mathbf{S}h\|_x^2 + \frac{L_{\mathbf{S}}}{6}\|\mathbf{S}h\|_x^3$

[(r)] Works Polyak (2009), Mishchenko (2021), Doikov and Nesterov (2021) have explicit updates and iteration cost $\mathcal{O}(d^3)$, but for the costs of slower global rate, slower local rate, and slower local rate, respectively.

[(lin)] (Hanzely et al., 2022) didn't show global linear rate of AICN. However, it follows from our Theorems 4, 3 for $\mathbf{S}_k = \mathbf{I}$.

Lemma 3 (Contractive properties of projection matrix $\mathbf{P}_x^{\mathbf{S}}$). *For any $g, h \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^2\right] = h^\top \nabla^2 f(x) \mathbb{E}\left[\mathbf{P}_x^{\mathbf{S}}\right] h \stackrel{As.1}{=} \frac{\tau}{d}\|h\|_x^2, \tag{21}$$

$$\mathbb{E}\left[\left\|\mathbf{P}_x^{\mathbf{S}} g\right\|_x^{*2}\right] = g^\top \mathbb{E}\left[\mathbf{P}_x^{\mathbf{S}}\right] [\nabla^2 f(x)]^\dagger g \stackrel{As.1}{=} \frac{\tau}{d}\|g\|_x^{*2}, \tag{22}$$

$$\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^2 \leq \left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^2 + \left\|(\mathbf{I} - \mathbf{P}_x^{\mathbf{S}}) h\right\|_x^2 = \|h\|_x^2, \tag{23}$$

$$\mathbb{E}\left[\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^3\right] \leq \mathbb{E}\left[\|h\|_x \cdot \left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^2\right] = \|h\|_x \mathbb{E}\left[\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_x^2\right] \stackrel{As.1}{=} \frac{\tau}{d}\|h\|_x^3. \tag{24}$$

We can bound condition number of the expected projection matrix. Define

$$\hat{\mathbf{P}_x^{\mathbf{S}}} \stackrel{\text{def}}{=} [\nabla^2 f(x)]^{\frac{1}{2}} \mathbf{S} \left[\nabla_{\mathbf{S}}^2 f(x)\right]^\dagger \mathbf{S}^\top [\nabla^2 f(x)]^{\frac{1}{2}} = [\nabla^2 f(x)]^{\frac{1}{2}} \mathbf{P}_x^{\mathbf{S}} [\nabla^2 f(x)]^{\dagger \frac{1}{2}}.$$

Proposition 2 (Analogy to Lemma 7 in (Gower et al., 2019)). *For $\mathbf{S} \sim \mathcal{D}$ satisfying conditions*

$$\textit{Null}(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}) = \textit{Null}(\mathbf{S}) \qquad \textit{and} \qquad \textit{Range}(\nabla^2 f(x)) \subset \textit{Range}\left(\mathbb{E}\left[\mathbf{S}_k \mathbf{S}_k^\top\right]\right), \tag{25}$$

also exactness condition holds

$$\textit{Range}(\nabla^2 f(x)) = \textit{Range}\left(\mathbb{E}\left[\hat{\mathbf{P}_x^{\mathbf{S}}}\right]\right), \tag{26}$$

and formula for $\rho(x)$ can be simplified

$$\rho(x) = \lambda_{\textit{min}}^+\left(\mathbb{E}\left[\alpha_{x,\mathbf{S}} \mathbf{P}_x^{\mathbf{S}}\right]\right) > 0 \tag{27}$$

and bounded $0 < \rho(x) \leq 1$. *Consequently,* $0 < \rho \leq 1$.

B.2. Function upperbound as Regularized Newton method

We can show a key idea from Regularized Newton methods: that $T_{\mathbf{S}}(x, h)$ is the function value upper bound, and minimizing it in h decreases the function value.

Proposition 3 (Lemma 2 in (Hanzely et al., 2022))**.** *For L_{semi}-semi-strong self-concordant f, and any $x \in \mathbb{R}^d, h \in \mathbb{R}^{\tau(\mathbf{S})}$, sketches $\mathbf{S} \in \mathbb{R}^{d \times \tau(\mathbf{S})}$ and $x_+ \overset{def}{=} x + \mathbf{S}h$ it holds*

$$\left| f(x_+) - f(x) - \langle \nabla f(x), \mathbf{S}h \rangle - \tfrac{1}{2} \|\mathbf{S}h\|_x^2 \right| \leq \tfrac{L_{semi}}{6} \|\mathbf{S}h\|_x^3, \tag{28}$$

$$f(x_+) \leq T_{\mathbf{S}}(x, h), \tag{29}$$

hence for $h^* \overset{def}{=} \operatorname{argmin}_{h \in \mathbb{R}^{\tau(\mathbf{S})}} T_{\mathbf{S}}(x, h)$ and corresponding x_+ we have functional value decrease,

$$f(x_+) \leq T_{\mathbf{S}}(x, h^*) = \min_{h \in \tau(\mathbf{S})} T_{\mathbf{S}}(x, h) \leq T_{\mathbf{S}}(x, 0) = f(x).$$

Next we show one step decrease in local sketched norms.

Lemma 4. *For $L_{\mathbf{S}}$-self-concordant function f, updates SGN, (5) decrease functional value as*

$$f(x_k) - f(x_{k+1}) \geq \left(2 \max \left\{ \sqrt{L_{est} \|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k, \mathbf{S}_k}^*}, 2 \right\} \right)^{-1} \|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k, \mathbf{S}_k}^{*2}. \tag{30}$$

We can show show one step decrease based on a virtual point y. Following lemma is crucial for global convex convergence.

Lemma 5. *Fix any $y \in \mathbb{R}^d$. Let the function f be L_{semi}-semi-strong self-concordant and sketch matrices $\mathbf{S}_k \sim \mathcal{D}$ have unbiased projection matrix, Assumption 1. Then SGN has decrease*

$$\mathbb{E}\left[f(x_{k+1}) | x_k \right] \leq \left(1 - \tfrac{\tau}{d} \right) f(x_k) + \tfrac{\tau}{d} f(y) + \tfrac{\tau}{d} \tfrac{\max L_{est} + L_{semi}}{6} \|y - x_k\|_{x_k}^3. \tag{31}$$

B.3. Convergence limitations of sketch-and-project methods

Similarly to AICN, we can show a quadratic decrease of the gradient norm in the sketched direction.

Lemma 6. *For L_{semi}-semi-strong self-concordant function f and parameter choice $L_{est} \geq L_{semi}$, one step of SGN has quadratic decrease in the Range$(\mathbf{S})$,*

$$\|\nabla_{\mathbf{S}} f(x_{k+1})\|_{x_k, \mathbf{S}}^* \leq L_{est} \alpha_{k, \mathbf{S}_k}^2 \|\nabla_{\mathbf{S}} f(x_k)\|_{x_k, \mathbf{S}}^{*2}. \tag{32}$$

Nevertheless, this is insufficient for superlinear local convergence; we can achieve a linear rate at best. We illustrate this on an edge case where f is a quadratic function. Then self-concordance holds with $L_{\mathbf{S}} = 0$ and as $\alpha_{k, \mathbf{S}_k} \xrightarrow{L_{\mathbf{S}} \to 0} 1$, SGN stepsize becomes 1 and SGN simplifies to subspace Newton method. Unfortunately, it has just linear local convergence (Gower et al., 2019).

B.4. Why is global linear convergence achievable?

Gower et al. (2019) shows that updates $x_+ = x + \mathbf{S}h$, where h is a minimizer of RHS of (13) converge linearly and can be written as Newton method with stepsize $\frac{1}{\hat{L}}$. Conversely, our stepsize $\alpha_{k, \mathbf{S}_k}$ varies (9), so it is not directly applicable to us. However, a small tweak will do the trick. Observe following:

- We already have fast local convergence (Theorem 3), so we just need to show linear convergence for points $\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k, \mathbf{S}_k}^* \geq \frac{4}{L_{\mathbf{S}_k}}$.

- For bounded stepsize $\alpha_{k, \mathbf{S}_k}$ smaller than $\frac{1}{\hat{L}}$, we can follow global linear proof of RSN.

- Stepsize $\alpha_{k, \mathbf{S}_k}$ of SGN, (9), is inversely proportional to $L_{\text{est}} \|\nabla_{\mathbf{S}_k} f(x)\|_x^*$. Increasing L_{est} decreases the convergence neighborhood arbitrarily. We just need to express this in terms of L_{est}.

• From regularized Newton method perspective (4), we have

$$x_+ = x + \mathbf{S} \operatorname*{argmin}_{h \in \mathbb{R}^{\tau(\mathbf{S})}} \left(f(x) + \langle \nabla_{\mathbf{S}} f(x), h \rangle + \tfrac{1}{2} \left(1 + \tfrac{L_{\text{est}}}{3} \|h\|_{x,\mathbf{S}} \right) \|h\|_{x,\mathbf{S}}^2 \right),$$

hence if $1 + \frac{L_{\text{est}}}{3} \|h\|_{x,\mathbf{S}} \geq \hat{L}_{\mathbf{S}}$, then (8) upperbounds on RHS of (13), and hence next iterate of SGN really minimizes function upperbound. Denote $\alpha_{x,\mathbf{S}}$ SGN stepsize in point x in range of $\mathbf{S}$. We express L_{est} as

$$1 + \tfrac{L_{\text{est}}}{3} \|h\|_{x,\mathbf{S}} \geq \hat{L}_{\mathbf{S}} \Leftrightarrow L_{\text{est}} \geq \frac{3(\hat{L}_{\mathbf{S}}-1)}{\alpha_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*} \Leftrightarrow 1 \geq \frac{3(\hat{L}_{\mathbf{S}}-1)}{-1+\sqrt{1+2L_{\text{est}}\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*}} \tag{33}$$

$$\Leftrightarrow L_{\text{est}} \geq \frac{3}{2} \frac{(\hat{L}_{\mathbf{S}}-1)(3\hat{L}_{\mathbf{S}}-1)}{\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*}. \tag{34}$$

And for $L_{\text{est}} \geq \sup_{\mathbf{S}} \frac{9}{8} L_{\mathbf{S}} \hat{L}_{\mathbf{S}}^2 > \sup_{\mathbf{S}} \frac{3}{8} L_{\mathbf{S}} (\hat{L}_{\mathbf{S}} - 1)(3\hat{L}_{\mathbf{S}} - 1)$ it holds while $\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^* \geq \frac{4}{L_{\mathbf{S}}}$.

B.5. Algorithm comparisons

For readers convenience, we include pseudocodes of the most relevant baseline algorithms: Exact Newton Descent (Algorithm 2), RSN (Algorithm 3), SSCN (Algorithm 4), AICN (Algorithm 5).

We include extended version of Table 1 in Table 5.

Algorithm 2 Exact Newton Descent (Karimireddy et al., 2018)

Requires: Initial point $x_0 \in \mathbb{R}^d$, c-stability bound $\sigma > c > 0$
for $k = 0, 1, 2 \dots$ **do**
  $x_{k+1} = x_k - \frac{1}{\sigma} \left[\nabla^2 f(x_k)\right]^\dagger \nabla f(x_k)$
end for

Algorithm 3 Randomized Subspace Newton (Gower et al., 2019)

Requires: Initial point $x_0 \in \mathbb{R}^d$, distribution of sketches $\mathcal{D}$, relative smoothness constant $L_{\text{rel}} > 0$
for $k = 0, 1, 2 \dots$ **do**
  Sample $\mathbf{S}_k \sim \mathcal{D}$
  $x_{k+1} = x_k - \frac{1}{L_{\text{rel}}} \mathbf{S}_k \left[\nabla_{\mathbf{S}_k}^2 f(x_k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x_k)$
end for

Algorithm 4 SSCN: Stochastic Subspace Cubic Newton (Hanzely et al., 2020)

Requires: Initial point $x_0 \in \mathbb{R}^d$, distribution of random matrices $\mathcal{D}$, Lipschitzness of Hessian constant $L_{\mathbf{S}} > 0$
for $k = 0, 1, 2 \dots$ **do**
  Sample $\mathbf{S}_k \sim \mathcal{D}$
  $x_{k+1} = x_k - \mathbf{S}_k \operatorname{argmin}_{h \in \mathbb{R}^d} \hat{T}_{\mathbf{S}_k}(x_k, h)$[a]
end for

[a] for $\hat{T}_{\mathbf{S}}(x, h) = \langle \nabla f(x), \mathbf{S}h \rangle + \frac{1}{2} \|\mathbf{S}h\|_x^2 + \frac{L_{\mathbf{S}}}{6} \|\mathbf{S}h\|_2^3$

Algorithm 5 Affine-Invariant Cubic Newton (Hanzely et al., 2022)

Requires: Initial point $x_0 \in \mathbb{R}^d$, estimate of semi-strong self-concordance $L_{\text{est}} \geq L_{\text{semi}} > 0$
for $k = 0, 1, 2 \dots$ **do**
  $\alpha_k = \frac{-1+\sqrt{1+2L_{\text{est}}\|\nabla f(x_k)\|_{x_k}^*}}{L_{\text{est}}\|\nabla f(x_k)\|_{x_k}^*}$
  $x_{k+1} = x_k - \alpha_k \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$[a]
end for

[a] Equival., $x_{k+1} = x_k - \operatorname{argmin}_{h \in \mathbb{R}^d} T(x_k, h)$, for $T(x, h) \stackrel{\text{def}}{=} \langle \nabla f(x), h \rangle + \frac{1}{2} \|h\|_x^2 + \frac{L_{\text{est}}}{6} \|h\|_x^3$

C. Experiments

We support our theory by comparing SGN to SSCN on logistic regression empirical risk minimization,

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \log \left(1 - e^{-b_i a_i^\top x} \right) + \frac{\mu}{2} \|x\|_2^2 \right\},$$

for data from LIBSVM (Chang and Lin, 2011), with features $\{(a_i, b_i)\}_{i=1}^m$ and labels $b_i \in \{-1, 1\}$.

To match practical considerations of SSCN and for the sake of simplicity, we adjust SGN in unfavorable way: **i)** we choose sketching matrices $\mathbf{S}$ to be unbiased in l_2 norms (instead of local hessian norms $\|\cdot\|_x$ from Assumption 1). **ii)** To disregard implementation specifics, we report iterations on the x-axis. Note that SSCN needs to use a subsolver (extra line-search) to

Table 5. Global convergence rate of low-rank Newton methods for convex and Lipschitz smooth functions. We use fastest full-dimensional algorithms as the baseline (for extended version, see Section 4). For simplicity, we disregard differences between various notions of smoothness.

Update direction / **Update oracle**	**Full-dimensional** (direction is deterministic)	**Low-rank** (direction in expectation)
Non-Newton direction	$\mathcal{O}(\mathbf{k^{-2}})$ Cubic Newton (Nesterov and Polyak, 2006) Glob. Regularized Newton (Mishchenko, 2021) (Doikov and Nesterov, 2021) $\mathcal{O}(k^{-\frac{1}{4}})$ Globally Regularized Newton (Polyak, 2009)	$\mathcal{O}(k^{-1})$ Stoch. Subspace Cubic Newton (Hanzely et al., 2020)
Newton direction	$\mathcal{O}(\mathbf{k^{-2}})$ Affine-Invariant Cubic Newton (Hanzely et al., 2022) $\mathcal{O}(k^{-1})$ Exact Newton Descent (Karimireddy et al., 2018) $\mathcal{O}(k^{-\frac{1}{2}})$ Damped Newton B (Nesterov and Nemirovski, 1994)	$\mathcal{O}(\mathbf{k^{-2}})$ **Sketchy Global Newton (this work)** $\mathcal{O}(k^{-1})$ Randomized Subspace Newton (Gower et al., 2019)

solve implicit step in each iteration. If naively implemented using matrix inverses, iterations of SSCN are $\times \log \frac{1}{\varepsilon}$ slower. We chose to didn't report time as this would naturally ask for optimized implementations and experiments on a larger scale – this was out of the scope of the paper. Figure 1 shows that despite simplicity of SGN and unfavourable adjustments, SGN performs comparably to SSCN.

In Figure 2 we include comparison of SGN and Accelerated Coordinate Descent on small-scale experiments.

C.1. Implementation

We use comparison framework from (Hanzely et al., 2020), including implementations of SSCN, Coordinate Descent and Accelerated Coordinate Descent.

Experiments are implemented in Python 3.6.9 and run on workstation with 48 CPUs Intel(R) Xeon(R) Gold 6246 CPU @ 3.30GHz. Total training time was less than 10 hours. Source code and instructions are included in supplementary materials. As we fixed random seed, experiments should be fully reproducible.

C.2. Insights from other papers contributions

We would also like to point out other properties of SGN based on experiments in related literature:

• **Rank of $\mathbf{S}$ and first-order methods:** Gower et al. (2019) showed a detailed comparison of the effect of various ranks of $\mathbf{S}$. Also, Gower et al. (2019) showed that RSN (fixed-stepsize Newton) is much faster than first-order Accelerated Gradient Descent for highly dense problems. For extremely sparse problems, Accelerated Gradient Descent has competitive performance. As the stepsize of SGN is increasing while getting close to the solution, we expect similar, if not better results.

• **Various sketch distributions**: Hanzely et al. (2020) considered various distributions of sketch matrices $\mathbf{S} \sim \mathcal{D}$. In all of their examples, SSCN outperformed CD with uniform or importance sampling and was competitive with Accelerated Gradient Descent. As SGN is competitive to SSCN, similar results should hold for SGN as well.

• **Local norms vs l_2 norms:** Hanzely et al. (2022) shows that the optimized implementation of AICN saves time in each iteration over the optimized implementation of Cubic Newton. As SGN and SSCN use the same updates (but in subspaces), it indicates that SGN saves time over SSCN.

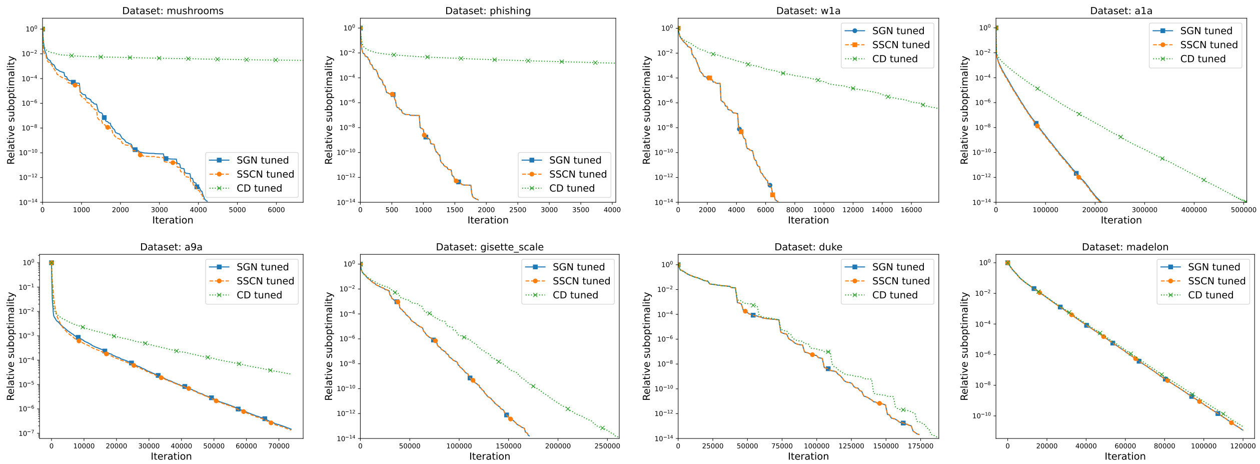

Figure 1. Comparison of SSCN, SGN and Coordinate Descent on logistic regression on LIBVSM datasets for sketch matrices **S** of rank one. We fine-tune all algorithms for smoothness parameters.

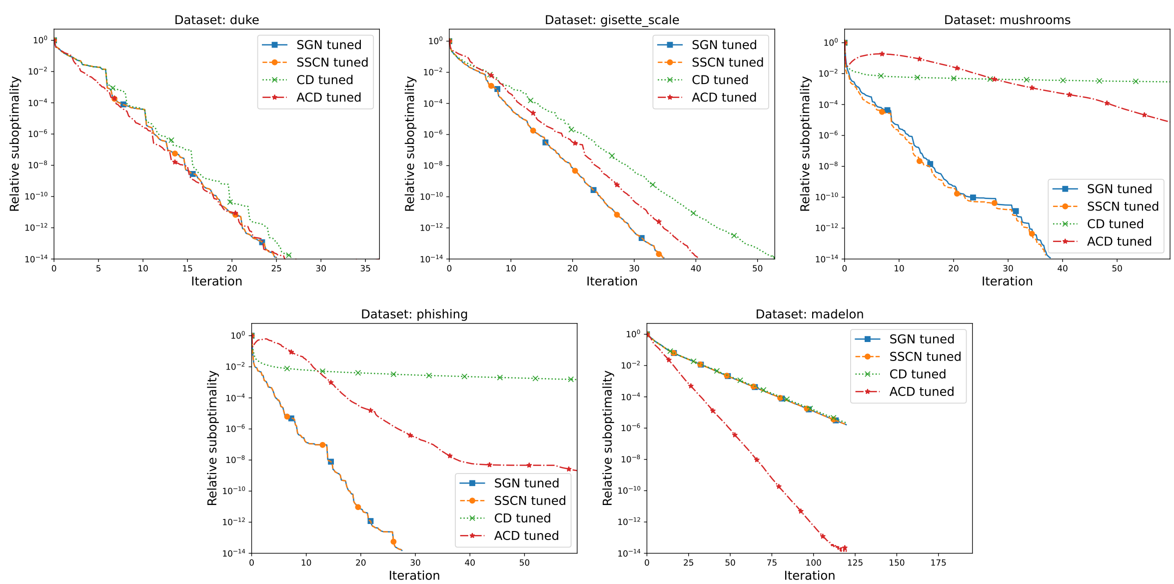

Figure 2. Comparison of SSCN, SGN, Coordinate Descent and Accelerated Coordinate Descent on logistic regression on LIBVSM datasets for sketch matrices **S** of rank one. We fine-tune all algorithms for smoothness parameters.

D. Proofs

For easier reference, we split proofs into four sections, based on the result category:

- Proofs explaining general properties of SGN (Appendix D.1)
- Proofs of global rate $\mathcal{O}\left(k^{-2}\right)$ in convex setup (Appendix D.2)
- Proofs of local linear rate (Appendix D.3)
- Proofs of global linear rate (Appendix D.4)

D.1. General properties

Proof of Theorem 1. (Three viewpoints of SGN)
Because $\nabla f(x_k) \in \text{Range}(\nabla^2 f(x_k))$, it holds $\nabla^2 f(x)[\nabla^2 f(x_k)]^\dagger \nabla f(x_k) = \nabla f(x_k)$. Updates (5) and (6) are equivalent as

$$\begin{aligned}\mathbf{P}_{x_k}^{\mathbf{S}_k}[\nabla^2 f(x_k)]^\dagger \nabla f(x_k) &= \mathbf{S}_k \left(\mathbf{S}_k^\top \nabla^2 f(x_k)\mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top \nabla^2 f(x_k)[\nabla^2 f(x_k)]^\dagger \nabla f(x_k)\\ &= \mathbf{S}_k \left(\mathbf{S}_k^\top \nabla^2 f(x_k)\mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top \nabla f(x_k)\\ &= \mathbf{S}_k[\nabla^2_{\mathbf{S}_k} f(x_k)]^\dagger \nabla_{\mathbf{S}_k} f(x_k)\end{aligned}$$

Taking gradient of $T_{\mathbf{S}_k}(x_k, h)$ w.r.t. h and setting it to 0 yields that for solution h^* holds

$$\nabla_{\mathbf{S}_k} f(x_k) + \nabla^2_{\mathbf{S}_k} f(x_k) h^* + \frac{L_{\text{est}}}{2}\|h^*\|_{x_k,\mathbf{S}_k} \nabla^2_{\mathbf{S}_k} f(x_k) h^* = 0 \tag{35}$$

which after rearranging is

$$h^* = -\left(1 + \frac{L_{\text{est}}}{2}\|h^*\|_{x_k,\mathbf{S}_k}\right)^{-1} \left[\nabla^2_{\mathbf{S}_k} f(x_k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x_k), \tag{36}$$

thus solution of cubical regulaziration in local norms (8) has form of Newton method with stepsize $\alpha_{k,\mathbf{S}_k} = \left(1 + \frac{L_{\text{est}}}{2}\|h^*\|_{x_k,\mathbf{S}_k}\right)^{-1}$. We are left to show that this $\alpha_{k,\mathbf{S}_k}$ is equivalent to (9).

Substitute h^* from (36) to (35) and $\alpha_{k,\mathbf{S}_k} = \left(1 + \frac{L_{\text{est}}}{2}\|h^*\|_{x_k,\mathbf{S}_k}\right)^{-1}$ and then use $\nabla^2 f(x)[\nabla^2 f(x_k)]^\dagger \nabla f(x_k) = \nabla f(x_k)$, to get

$$0 = \nabla_{\mathbf{S}_k} f(x_k) + \nabla^2_{\mathbf{S}_k} f(x_k)\left(-\alpha_{k,\mathbf{S}_k}\left[\nabla^2_{\mathbf{S}_k} f(x_k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x_k)\right) \tag{37}$$

$$+ \frac{L_{\text{est}}}{2}\left(\alpha_{k,\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}\right)\nabla^2_{\mathbf{S}_k} f(x_k)\left(-\alpha_{k,\mathbf{S}_k}\left[\nabla^2_{\mathbf{S}_k} f(x_k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x_k)\right) \tag{38}$$

$$= \left(1 - \alpha_{k,\mathbf{S}_k} - \frac{L_{\text{est}}}{2}\alpha^2_{k,\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}\right)\nabla_{\mathbf{S}_k} f(x_k). \tag{39}$$

Finally, $\alpha_{k,\mathbf{S}_k}$ from (9) is a positive root of polynomial $1 - \alpha_{k,\mathbf{S}_k} - \frac{L_{\text{est}}}{2}\alpha^2_{k,\mathbf{S}_k} = 0$, which concludes the equivalence of (5), (6) and (4). □

Lemma 7 (Stepsize bound)**.** *Stepsize $\alpha_{k,\mathbf{S}_k}$ can be bounded as*

$$\alpha_{k,\mathbf{S}_k} \leq \frac{\sqrt{2}}{\sqrt{L_{est}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}}, \tag{40}$$

and for x_k far from solution, $\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \geq \frac{4}{L_{\mathbf{S}_k}}$ and $L_{est} = \frac{9}{8}\sup_{\mathbf{S}} L_{\mathbf{S}}\hat{L}^2_{\mathbf{S}}$ holds $\alpha_{k,\mathbf{S}_k}\hat{L}_{\mathbf{S}_k} \leq \frac{2}{3}$.

Lemma 8 ((Gower et al., 2020))**.** *Matrix $\mathbf{P}_x^{\mathbf{S}}$ is a projection matrix on Range($\mathbf{S}$) w.r.t. norm $\|\cdot\|_x$.*

Proof of Lemma 8. (Matrix $\mathbf{P}_x^{\mathbf{S}}$ is a projection matrix)
For arbitrary square matrix $\mathbf{M}$ pseudoinverse guarantee $\mathbf{M}^\dagger \mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger$. Applying this to $M \leftarrow \left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}\right)$ yields $\left\langle \mathbf{P}_x^{\mathbf{S}} y, \mathbf{P}_x^{\mathbf{S}} z \right\rangle_{\nabla^2 f(x)} = \left\langle \mathbf{P}_x^{\mathbf{S}} y, z \right\rangle_{\nabla^2 f(x)}$ $y, z \in \mathbb{R}^d$. Thus, $\mathbf{P}_x^{\mathbf{S}}$ is really projection matrix w.r.t. $\|\cdot\|_x$. □

Proof of Lemma 1. (Unbiased $\mathbf{P}_x^{\mathbf{S}}$ implies $\mathbb{E}\left[\tau(\mathbf{S})\right] = \tau$, as in Lemma 5.2 of (Hanzely et al., 2020)) We use definitions and cyclic property of the matrix trace,

$$\mathbb{E}\left[\tau(\mathbf{S})\right] = \mathbb{E}\left[\operatorname{Tr}\left(\mathbf{I}^{\tau(\mathbf{S})}\right)\right] = \mathbb{E}\left[\operatorname{Tr}\left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S} \left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}\right)^\dagger\right)\right] = \mathbb{E}\left[\operatorname{Tr}\left(\mathbf{P}_x^{\mathbf{S}}\right)\right] \tag{41}$$

$$= \operatorname{Tr}\left(\frac{\tau}{d}\mathbf{I}^d\right) = \tau. \tag{42}$$

□

Proof of Lemma 2. (Construction of unbiased sketch matrices in local norms from ones in l_2 norms)
We have

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_x^{\mathbf{S}}\right] = \left[\nabla^2 f(x)\right]^{-1/2} \mathbb{E}_{\mathbf{M}\sim\tilde{\mathcal{D}}}\left[\mathbf{M}^\top \left(\mathbf{M}^\top \mathbf{M}\right)^\dagger \mathbf{M}\right] \left[\nabla^2 f(x)\right]^{1/2} \tag{43}$$

$$= \left[\nabla^2 f(x)\right]^{-1/2} \frac{\tau}{d}\mathbf{I} \left[\nabla^2 f(x)\right]^{1/2} = \frac{\tau}{d}\mathbf{I}. \tag{44}$$

□

Note that

$$h_k \overset{(5)}{=} -\alpha_{k,\mathbf{S}_k} \mathbf{S}_k [\nabla^2_{\mathbf{S}_k} f(x_k)]^\dagger \nabla_{\mathbf{S}_k} f(x_k), \qquad \|h_k\|_{x_k} = \alpha_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}. \tag{45}$$

Proof of Lemma 4. (One step functional value decrease in terms of norms of gradients)
For $h_k = x_{k+1} - x_k$, we can follow proof of Lemma 10 in Hanzely et al. (2022),

$$f(x_k) - f(x_{k+1}) \overset{(28)}{\geq} -\langle \nabla_{\mathbf{S}} f(x_k), h_k \rangle - \frac{1}{2}\|h_k\|^2_{x_k,\mathbf{S}_k} - \frac{L_{\text{est}}}{6}\|h\|^3_{x_k,\mathbf{S}_k} \tag{46}$$

$$\overset{(45)}{=} \alpha_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} - \frac{1}{2}\alpha^2_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} \tag{47}$$

$$- \frac{L_{\text{est}}}{6}\alpha^3_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*3}_{x_k,\mathbf{S}_k,\mathbf{S}} \tag{48}$$

$$= \left(1 - \frac{1}{2}\alpha_{k,\mathbf{S}_k} - \frac{L_{\text{est}}}{6}\alpha^2_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}\right) \alpha_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} \tag{49}$$

$$\geq \frac{1}{2}\alpha_{k,\mathbf{S}_k} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} \tag{50}$$

$$\geq \frac{1}{2\max\left\{\sqrt{L_{\text{est}} \|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}, 2\right\}} \|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k}. \tag{51}$$

□

Proof of Lemma 6. (Quadratic local decrease in subspaces).
We bound norm of $\nabla_{\mathbf{S}} f(x_{k+1})$ using basic norm manipulation and triangle inequality as

$$\begin{aligned}
&\|\nabla_{\mathbf{S}_k} f(x_{k+1})\|^*_{x_k,\mathbf{S}_k} \\
&= \left\|\nabla_{\mathbf{S}_k} f(x_{k+1}) - \nabla^2_{\mathbf{S}_k} f(x_k)(x_{k+1} - x_k) - \alpha_{k,\mathbf{S}_k} \nabla_{\mathbf{S}_k} f(x_k)\right\|^*_{x_k,\mathbf{S}_k} \\
&= \left\|\nabla_{\mathbf{S}_k} f(x_{k+1}) - \nabla_{\mathbf{S}_k} f(x_k) - \nabla^2_{\mathbf{S}_k} f(x_k)(x_{k+1} - x_k) + (1 - \alpha_{k,\mathbf{S}_k}) \nabla_{\mathbf{S}_k} f(x_k)\right\|^*_{x_k,\mathbf{S}_k} \\
&\leq \left\|\nabla_{\mathbf{S}_k} f(x_{k+1}) - \nabla_{\mathbf{S}_k} f(x_k) - \nabla^2_{\mathbf{S}_k} f(x_k)(x_{k+1} - x_k)\right\|^*_{x_k,\mathbf{S}_k} + (1 - \alpha_{k,\mathbf{S}_k}) \|\nabla f(x_k)\|^*_{x_k,\mathbf{S}_k}
\end{aligned}$$

Using L_{semi}-semi-strong self-concordance, we can continue

$$
\begin{aligned}
&\leq \left\|\nabla_{\mathbf{S}_k} f(x_{k+1}) - \nabla_{\mathbf{S}_k} f(x_k) - \nabla^2_{\mathbf{S}_k} f(x_k)(x_{k+1}-x_k)\right\|^*_{x_k,\mathbf{S}_k} + (1-\alpha_{k,\mathbf{S}_k})\|\nabla_{\mathbf{S}} f(x_k)\|^*_{x_k,\mathbf{S}_k} \\
&\leq \frac{L_{\text{semi}}}{2}\|x_{k+1}-x_k\|^2_{x_k,\mathbf{S}_k} + (1-\alpha_{k,\mathbf{S}_k})\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \\
&= \frac{L_{\text{semi}}\alpha^2_{k,\mathbf{S}_k}}{2}\|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} + (1-\alpha_{k,\mathbf{S}_k})\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}} \\
&\leq \frac{L_{\text{est}}\alpha^2_{k,\mathbf{S}_k}}{2}\|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k} + (1-\alpha_{k,\mathbf{S}_k})\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \\
&= \left(\frac{L_{\text{est}}\alpha^2_{k,\mathbf{S}_k}}{2}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}} - \alpha_{k,\mathbf{S}_k} + 1\right)\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \\
&\overset{(9)}{=} L_{\text{est}}\alpha^2_{k,\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|^{*2}_{x_k,\mathbf{S}_k}.
\end{aligned}
$$

Last equality holds because of the choice of $\alpha_{k,\mathbf{S}_k}$. □

D.1.1. TECHNICAL LEMMAS

Lemma 9 (Arithmetic mean – Geometric mean inequality)**.** *For $c \geq 0$ we have*

$$1 + c = \frac{1 + (1+2c)}{2} \overset{AG}{\geq} \sqrt{1+2c}. \tag{52}$$

Lemma 10 (Jensen for square root)**.** *Function $f(x) = \sqrt{x}$ is concave, hence for $c \geq 0$ we have*

$$\frac{1}{\sqrt{2}}(\sqrt{c}+1) \leq \sqrt{c+1} \qquad \leq \sqrt{c}+1. \tag{53}$$

Proof of Lemma 7. Denote $G_k \overset{\text{def}}{=} L_{\text{est}}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}$. Using (53) with $c \leftarrow 2G > 0$ and

$$\alpha_{k,\mathbf{S}_k} = \frac{-1+\sqrt{1+2G}}{G} \leq \frac{\sqrt{2G}}{G} = \frac{\sqrt{2}}{\sqrt{G}} = \frac{\sqrt{2}}{\sqrt{L_{\text{est}}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}} \tag{54}$$

and

$$
\begin{aligned}
\alpha_{k,\mathbf{S}_k}\hat{L}_{\mathbf{S}_k} &\leq \frac{\sqrt{2}\hat{L}_{\mathbf{S}_k}}{\sqrt{L_{\text{est}}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}} && (55)\\
&\leq \frac{\sqrt{2}\hat{L}_{\mathbf{S}_k}}{\sqrt{\frac{9}{8}L_{\mathbf{S}_k}\hat{L}^2_{\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}} && (56)\\
&\leq \frac{4}{3}\frac{1}{\sqrt{L_{\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k}}} \leq \frac{2}{3} \qquad \text{for } \|\nabla_{\mathbf{S}_k} f(x_k)\|^*_{x_k,\mathbf{S}_k} \geq \frac{4}{\hat{L}_{\mathbf{S}_k}}. && (57)
\end{aligned}
$$

□

D.2. Global convex rate

Proof of Lemma 5. (Key lemma for global convex convergence). Denote

$$\Omega_{\mathbf{S}}(x,h') \overset{\text{def}}{=} f(x)\left\langle \nabla f(x), \mathbf{P}^{\mathbf{S}}_x h'\right\rangle + \frac{1}{2}\left\|\mathbf{P}^{\mathbf{S}}_x h'\right\|^2_x + \frac{L_{\text{est}}}{6}\left\|\mathbf{P}^{\mathbf{S}}_x h'\right\|^3_x, \tag{58}$$

so that

$$\min_{h'\in\mathbb{R}^d}\Omega_{\mathbf{S}}(x,h') = \min_{h\in\mathbb{R}^{\tau(\mathbf{S})}} T_{\mathbf{S}}(x,h). \tag{59}$$

For arbitrary $y \in \mathbb{R}^d$ denote $h \overset{\text{def}}{=} y - x_k$. We can calculate

$$
\begin{aligned}
f(x_{k+1}) &\leq \min_{h' \in \mathbb{R}^{\tau(\mathbf{S})}} T_{\mathbf{S}}(x_k, h') = \min_{h'' \in \mathbb{R}^d} \Omega_{\mathbf{S}}(x_k, h'') && (60)\\
\mathbb{E}\left[f(x_{k+1})\right] &\leq \mathbb{E}\left[\Omega_{\mathbf{S}}(x_k, h)\right] && (61)\\
&= f(x_k) + \frac{\tau}{d}\left\langle \nabla f(x_k), h\right\rangle + \frac{1}{2}\mathbb{E}\left[\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_{x_k}^2\right] + \mathbb{E}\left[\frac{L_{\text{est}}}{6}\left\|\mathbf{P}_x^{\mathbf{S}} h\right\|_{x_k}^3\right] && (62)\\
&\overset{(21)}{\leq} f(x_k) + \frac{\tau}{d}\left\langle \nabla f(x_k), h\right\rangle + \frac{\tau}{2d}\|h\|_{x_k}^2 + \frac{L_{\text{est}}}{6}\frac{\tau}{d}\|h\|_{x_k}^3 && (63)\\
&\overset{(28)}{\leq} f(x_k) + \frac{\tau}{d}\left(f(y) - f(x_k) + \frac{L_{\text{semi}}}{6}\|y - x_k\|_{x_k}^3\right) + \frac{L_{\text{est}}}{6}\frac{\tau}{d}\|h\|_{x_k}^3, && (64)
\end{aligned}
$$

In second to last inequality depends on unbiasedness of projection $\mathbf{P}_x^{\mathbf{S}}$, Assumption 1. In last inequality we used semi-strong self-concordance, Proposition 3 with $\mathbf{S} = \mathbf{I}$. □

Proof of Theorem 2. (Global convex rate). Denote

$$
A_0 \overset{\text{def}}{=} \frac{4}{3}\left(\frac{d}{\tau}\right)^3, \tag{65}
$$

$$
A_k \overset{\text{def}}{=} A_0 + \sum_{t=1}^{k} t^2 = A_0 - 1 + \frac{k(k+1)(2k+1)}{6} \geq A_0 + \frac{k^3}{3}, \tag{66}
$$

$$
\text{...consequently} \qquad \sum_{t=1}^{k} \frac{t^6}{A_t^2} \leq 9k, \tag{67}
$$

$$
\eta_t \overset{\text{def}}{=} \frac{d}{\tau}\frac{(t+1)^2}{A_{t+1}} \qquad \text{implying } 1 - \frac{d}{\tau}\eta_t = \frac{A_t}{A_{t+1}}. \tag{68}
$$

Note that this choice of A_0 implies (Hanzely et al., 2020)

$$
\eta_{t-1} \leq \frac{d}{\tau}\frac{t^2}{A_0 + \frac{t^3}{3}} \leq \frac{d}{\tau}\sup_{t\in\mathbb{N}}\frac{t^2}{A_0 + \frac{t^3}{3}} \leq \frac{d}{\tau}\sup_{\zeta>0}\frac{\zeta^2}{A_0 + \frac{\zeta^3}{3}}, = 1 \tag{69}
$$

and $\eta_t \in [0, 1]$. Set $y \overset{\text{def}}{=} \eta_t x_* + (1 - \eta_t)x_t$ in Lemma 5. From convexity of f,

$$
\mathbb{E}\left[f(x_{t+1}|x_t)\right] \leq \left(1 - \frac{\tau}{d}\right)f(x_t) + \frac{\tau}{d}f_*\eta_t + \frac{\tau}{d}f(x_t)(1-\eta_t) + \frac{\tau}{d}\left(\frac{\max L_{\mathbf{S}} + L_{\text{semi}}}{6}\|x_t - x_*\|_{x_t}^3\eta_t^3\right). \tag{70}
$$

Denote $\delta_t \overset{\text{def}}{=} \mathbb{E}\left[f(x_t) - f_*\right]$. Subtracting f_* from both sides and substituting η_k yields

$$
\delta_{t+1} \leq \frac{A_t}{A_{t+1}}\delta_t + \frac{\max L_{\mathbf{S}} + L_{\text{semi}}}{6}\|x_t - x_*\|_{x_t}^3\left(\frac{d}{\tau}\right)^2\left(\frac{(t+1)^2}{A_{t+1}}\right)^3. \tag{71}
$$

Multiplying by A_{t+1} and summing from from $t = 0, \dots, k-1$ yields

$$
A_k\delta_k \leq A_0\delta_0 + \frac{\max L_{\mathbf{S}} + L_{\text{semi}}}{6}\frac{d^2}{\tau^2}\sum_{t=0}^{k-1}\|x_t - x_*\|_{x_t}^3\frac{(t+1)^6}{A_{t+1}^2}, \tag{72}
$$

Using $\sup_{x\in\mathcal{Q}(x_0)}\|x - x_*\|_x \leq R$ we can simplify and shift summation indices,

$$
\begin{aligned}
A_k\delta_k &\leq A_0\delta_0 + \frac{\max L_{\mathbf{S}} + L_{\text{semi}}}{6}\frac{d^2}{\tau^2}D^3\sum_{t=1}^{k}\frac{t^6}{A_t^2} && (73)\\
&\leq A_0\delta_0 + \frac{\max L_{\mathbf{S}} + L_{\text{semi}}}{6}\frac{d^2}{\tau^2}D^3 9k && (74)
\end{aligned}
$$

and

$$\delta_k \leq \frac{A_0\delta_0}{A_k} + \frac{3(\max L_{\mathbf{S}} + L_{\text{semi}})d^2D^3k}{2\tau^2 A_k} \tag{75}$$

$$\leq \frac{3A_0\delta_0}{k^3} + \frac{9(\max L_{\mathbf{S}} + L_{\text{semi}})d^2D^3}{2\tau^2 k^2} \tag{76}$$

which concludes the proof. □

D.3. Local linear rate

Proposition 4 (Lemma E.3 in Hanzely et al. (2020)). *For $\gamma > 0$ and x_k in neighborhood $x_k \in \left\{x : \|\nabla f(x)\|_x^* < \frac{2}{(1+\gamma^{-1})L_{sc}}\right\}$ for L_{sc}-self-concordant function f, we can bound*

$$f(x_k) - f_* \leq \frac{1}{2}(1+\gamma)\|\nabla f(x_k)\|_{x_k}^{*2}. \tag{77}$$

Proof of Theorem 3. (Fast local linear rate theorem).
Proposition 4 with $\gamma = 2$ implies that in neighborhood $\|\nabla f(x_k)\|_{x_k,\mathbf{S}}^{*2} \leq \frac{4}{L_{\mathbf{S}}}$,

$$f(x_k) - f(x_{k+1}) \overset{(30)}{\geq} \frac{1}{4}\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k,\mathbf{S}_k}^{*2}$$

and with identity $\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}_k}^{*2} = \left\|\mathbf{P}_x^{\mathbf{S}}\nabla f(x)\right\|_x^{*2}$, we can continue

$$\mathbb{E}\left[f(x_k) - f(x_{k+1})\right] \overset{(30)}{\geq} \mathbb{E}\left[\frac{1}{4}\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k,\mathbf{S}}^{*2}\right] = \mathbb{E}\left[\frac{1}{4}\left\|\mathbf{P}_{x_k}^{\mathbf{S}_k}\nabla f(x_k)\right\|_{x_k}^{*2}\right] \tag{78}$$

$$\overset{(22)}{=} \frac{\tau}{4d}\|\nabla f(x_k)\|_{x_k}^{*2} \overset{(77)}{\geq} \frac{\tau}{2d(1+\gamma)}(f(x_k) - f_*). \tag{79}$$

Hence

$$\mathbb{E}\left[f(x_{k+1}) - f_*)\right] \leq \left(1 - \frac{\tau}{2d(1+\gamma)}\right)(f(x_k) - f_*),$$

and to finish the proof, we use tower property across iterates $x_0, x_1, \dots, x_k$. □

D.4. Global linear rate

Proposition 5 ((47) in Gower et al. (2019)). *Relative convexity* (14) *implies following bound*

$$f_* \leq f(x_k) - \frac{1}{2\hat{\mu}}\|\nabla f(x_k)\|_{x_k}^{*2}. \tag{80}$$

Proof of Theorem 4. (Global linear convergence under relative convexity)
Replacing $x \leftarrow x_k$ and $h \leftarrow \alpha_{k,\mathbf{S}_k}\mathbf{P}_{x_k}^{\mathbf{S}_k}[\nabla^2 f(x_k)]^\dagger\nabla f(x_k)$ so that $x_{k+1} = x_k + \mathbf{S}h$ in (13) yields

$$f(x_{k+1}) \leq f(x_k) - \alpha_{k,\mathbf{S}_k}\left(1 - \frac{1}{2}\hat{L}_{\mathbf{S}_k}\alpha_{k,\mathbf{S}_k}\right)\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k,\mathbf{S}_k}^{*2} \tag{81}$$

$$\leq f(x_k) - \frac{2}{3}\alpha_{k,\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k,\mathbf{S}_k}^{*2}. \tag{82}$$

In last step, we used that $\hat{L}_{\mathbf{S}_k}\alpha_{k,\mathbf{S}_k} \leq \frac{2}{3}$ holds for $\|\nabla_{\mathbf{S}_k} f(x_k)\|_{x_k,\mathbf{S}_k}^{*} \geq \frac{4}{\hat{L}_{\mathbf{S}_k}}$ (Lemma 7). Next, we take expectation on x_k and use definition of $\rho(x_k)$.

$$\mathbb{E}\left[f(x_{k+1})\right] \leq f(x_k) - \frac{2}{3}\|\nabla f(x_k)\|^2_{\mathbb{E}\left[\alpha_{k,\mathbf{S}_k}\mathbf{S}\left[\nabla^2_{\mathbf{S}_k} f(x_k)\right]^\dagger\mathbf{S}^\top\right]} \tag{83}$$

$$\leq f(x_k) - \frac{2}{3}\rho(x_k)\|\nabla f(x_k)\|_{x_k}^{*2} \tag{84}$$

$$\overset{(80)}{\leq} f(x_k) - \frac{4}{3}\rho(x_k)\hat{\mu}\left(f(x_k) - f_*\right). \tag{85}$$

Now $\rho(x) \geq \rho$, and ρ is bounded in Proposition 2. Rearranging and subtracting f_* gives

$$\mathbb{E}\left[f(x_{k+1}) - f_*\right] \leq \left(1 - \frac{4}{3}\rho\hat{\mu}\right)(f(x_k) - f_*), \tag{86}$$

Which after towering across all iterates yields the statement. □