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

**Slavomír Hanzely**
Mohammed bin Zayed University of Artificial Intelligence

## Abstract

In this paper, we propose the first sketch-and-project Newton method with the fast $\mathcal{O}\left(k^{-2}\right)$ global convergence rate for self-concordant functions. Our method, SGN, can be viewed in three ways: i) as a sketch-and-project algorithm projecting updates of the Newton method, ii) as a cubically regularized Newton method in the sketched subspaces, and iii) as a damped Newton method in the sketched subspaces.

SGN inherits the best of all three worlds: the cheap iteration costs of the sketch-and-project methods, the state-of-the-art $\mathcal{O}\left(k^{-2}\right)$ global convergence rate of the full-rank Newton-like methods, and the algorithm simplicity of the damped Newton methods. Finally, we empirically show SGN performs on par with baseline algorithms.

## 1 INTRODUCTION

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

Second-order methods are highly desirable due to their invariance to rescalings 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, the 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, the Newton method, achieves an extremely fast, quadratic convergence rate (precision doubles in each iteration) when initialized sufficiently close to the solution. However, the convergence of the Newton method is limited only to the neighborhood of the solution. Jarre and Toint (2016), Mascarenhas (2007), Bolte and Pauwels (2022) show that the line search and the trust-region Newton-like methods can diverge when initialized for from the optimum, even for convex problems.

### 1.1 Demands of Modern Machine Learning

Despite its long history, research on second-order methods has been thriving to this day. Newton-like methods with the fast $\mathcal{O}\left(k^{-2}\right)$ global rate were introduced relatively recently under the names Cubic Newton method (Nesterov and Polyak, 2006) or Globally regularized Newton methods (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 and inexact approximations can be used to avoid computing the inverse Hessian, simply storing the Hessian is impractical when the dimensionality $d$ is large. This challenge has served as a catalyst for the recent developments. To address the curse of dimensionality, Pilanci and Wainwright (2017) proposed sketching the Hessian matrix while keeping the gradient intact. Furthermore, Qu et al. (2016), Luo et al. (2016), Gower et al. (2019), Doikov and Richtárik (2018), and Hanzely et al. (2020) proposed Newton-like methods

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

Table 1: Global convergence rates of low-rank Newton methods for convex and smooth functions. We report dependence on the number of iterations $k$. We use the fastest full-dimensional algorithms as the baseline, we highlight the best rate in blue.

| **Update direction** \ **Update oracle** | **Full-dimensional** (direction is deterministic) | **Low-rank** (direction in expectation) |
|---|---|---|
| **Non-Newton direction** | $\mathcal{O}(k^{-2})$<br>Cubically regularized Newton (Nesterov and Polyak, 2006),<br>Globally regularized Newton (Mishchenko, 2021; Doikov and Nesterov, 2023) | $\mathcal{O}(k^{-1})$<br>Stochastic Subspace Cubic Newton (Hanzely et al., 2020) |
| **Newton direction** | $\mathcal{O}(k^{-2})$<br>Affine-Invariant Cubic Newton (Hanzely et al., 2022) | $\mathcal{O}(k^{-2})$<br>**Sketchy Global Newton (new)**<br>$\mathcal{O}(k^{-1})$<br>Randomized Subspace Newton (Gower et al., 2019) |

operating in random low-dimensional subspaces. This approach, also known as sketch-and-project (Gower and Richtárik, 2015), substantially reduces the computational cost per iteration. However, this happens at 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 the sketch-and-project adaptations of second-order methods can be improved. To this end, we introduce the **first** sketch-and-project method (SGN, Algorithm 1) boasting a global $\mathcal{O}\left(k^{-2}\right)$ convex convergence rate, matching dependence on iteration number $k$ of full-dimensional regularized Newton methods. Surprisingly, sketching on 1-dimensional subspaces with an iteration cost of $\mathcal{O}(1)$ (Gower et al., 2019) engenders $\mathcal{O}\left(k^{-2}\right)$ global convex rate.

As a cherry on top, SGN offers additional benefits in the form of a local linear convergence rate independent of the condition number and a global linear rate under the assumption of relative convexity (Definition 3). We summarize the contributions below and in Tables 2, 3.

- **One connects all:** We present SGN through three orthogonal viewpoints: the sketch-and-project method, the subspace Newton method, and the subspace regularized Newton method. Compared to the established algorithms, SGN can be viewed as AICN (Hanzely et al., 2022) operating in subspaces, SSCN (Hanzely et al., 2020) operating in local norms, or RSN (Gower et al., 2019) with the new stepsize schedule (Table 2).

  Designing an algorithm that preserves all desired properties was a significant challenge. It required a multitude of insights from the literature and an extremely careful analysis technique. Even the smallest misalignment can cause significantly slower convergence rate guarantees (see Section 5.2).
- **Fast global convergence:** SGN is the **first low-rank method** that minimizes convex functions with $\mathcal{O}\left(k^{-2}\right)$ global convergence, matching full-rank Newton-like methods. Other sketch-and-project methods (e.g., SSCN and RSN), have slower $\mathcal{O}\left(k^{-1}\right)$ rate (Table 1).
- **Cheap iterations:** SGN uses $\tau$–dimensional updates. Per-iteration cost is $\mathcal{O}\left(d\tau^2\right)$ and in the $\tau = 1$ case it is even $\mathcal{O}\left(1\right)$ (Gower et al., 2019). Conversely, full-rank Newton-like methods have cost $\mathcal{O}(d^3)$.
- **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 with a different smoothness constant achieves global linear rate $\mathcal{O}\left(\frac{L_{\text{alg}}}{\rho\hat{\mu}}\log\frac{1}{\varepsilon}\right)$ to a neighborhood of the solution (Theorem 4)[1].
- **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, simplifying parameter tuning. Update rule (4) is simple and explicit. Conversely, most fast globally convergent Newton-like algorithms require solving an extra subproblem in each iteration.

[1] $\rho$ is condition number of a projection matrix (22), and $L_{\text{alg}}$ is upperbound on semi-strong self-concordance (Definition 1) affecting the stepsize (7).

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

- **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 removes geometric interpretability, leads to worse constants, and ultimately causes the slower $\mathcal{O}\left(k^{-1}\right)$ convergence rate.

### 1.3 Objective

In this chapter, we consider the optimization objective

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

where function $f:\mathbb{R}^d\to\mathbb{R}$ is convex, twice differentiable with positive definite Hessian, bounded from below, and potentially ill-conditioned. The number of features $d$ is large. Denote the solution $x^* \overset{\text{def}}{=} \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$ and $f^* \overset{\text{def}}{=} f(x^*)$. We solve objective (1) using subspace methods, with a sparse update

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

where $\mathbf{S}\in\mathbb{R}^{d\times\tau(\mathbf{S})}, \mathbf{S}\sim\mathcal{D}$ is a thin matrix of rank $\tau(\mathbf{S})$ and $h\in\mathbb{R}^{\tau(\mathbf{S})}$. We denote gradients and Hessians along the subspace spanned by columns of $\mathbf{S}$ as $\nabla_{\mathbf{S}} f(x) \overset{\text{def}}{=} \mathbf{S}^\top\nabla f(x)$ and $\nabla^2_{\mathbf{S}} f(x) \overset{\text{def}}{=} \mathbf{S}^\top\nabla^2 f(x)\mathbf{S}$. Gower et al. (2019) shows that $\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}$ can be obtained by twice differentiating function $\lambda\to f(x+\mathbf{S}\lambda)$ at cost of $\tau(\mathbf{S})$ times of evaluating function $f(x+\mathbf{S}\lambda)$ by using reverse accumulation techniques (Christianson, 1992; Gower and Mello, 2012). For $\mathbf{S}\sim\mathcal{D}$ with a constant rank $\tau \overset{\text{def}}{=} \tau(\mathbf{S})$, it requires $\mathcal{O}\left(d\tau^2\right)$ arithmetic operations and in the case $\tau=1$, the cost can be reduced to even $\mathcal{O}\left(1\right)$ (Gower et al., 2019).

### 1.4 Affine-Invarant Geometry

We can define norms based on a symmetric positive definite matrix $\mathbf{H}\in\mathbb{R}^{d\times d}$. For any $x,g\in\mathbb{R}^d$,

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

As a special case $\mathbf{H}=\mathbf{I}$, we get $l_2$ norm $\|x\|_{\mathbf{I}} = \langle x,x\rangle^{1/2}$. We will be setting $\mathbf{H}$ to be a Hessian at local point, $\mathbf{H}=\nabla^2 f(x)$, with the shorthand for $g,h\in\mathbb{R}^d$ as

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

The main advantage of the local Hessian norm $\|h\|_x$ is its affine-invariance, as the affine transformation $f(x)\to\phi(y) \overset{\text{def}}{=} f(\mathbf{A}y)$, and $x\to\mathbf{A}^{-1}y$ imply

$$\begin{aligned}\|z\|^2_{\nabla^2\phi(y)} &= \left\langle\nabla^2\phi(y)z,z\right\rangle = \left\langle\mathbf{A}^\top\nabla^2 f(\mathbf{A}y)\mathbf{A}z,z\right\rangle\\ &= \left\langle\nabla^2 f(x)h,h\right\rangle = \|h\|^2_{\nabla^2 f(x)}.\end{aligned}$$

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

**Requires:** Initial point $x^0\in\mathbb{R}^d$, distribution of sketch matrices $\mathcal{D}$ such that $\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_x\right]=\frac{\tau}{d}\mathbf{I}$, semi-strong self-concordance upper bound $L_{\text{alg}}\geq L_{\text{semi}}$
**for** $k=0,1,2\ldots$ **do**
  Sample $\mathbf{S}_k\sim\mathcal{D}$
  $\alpha_k = \frac{2}{1+\sqrt{1+2L_{\text{alg}}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k}}}$
  $x^{k+1} = x^k - \alpha_k\mathbf{S}_k\left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger\nabla_{\mathbf{S}_k} f(x^k)$
**end for**

On the other hand, induced norm $\|h\|_{\mathbf{I}}$ is not, because

$$\|z\|^2_{\mathbf{I}} = \langle z,z\rangle = \left\langle\mathbf{A}^{-1}h,\mathbf{A}^{-1}h\right\rangle = \left\|\mathbf{A}^{-1}h\right\|^2_{\mathbf{I}}.$$

With respect to geometry around point $x$, the more natural norm is the local Hessian norm, $\|h\|_{\nabla f(x)}$. Affine-invariance implies that its level sets $\left\{y\in\mathbb{R}^d \,|\, \|y-x\|^2_x\leq c\right\}$ are balls centered around $x$ (all directions have the same scaling). In comparison, the scaling of the $l_2$ norm is dependent on the eigenvalues of the Hessian. In terms of convergence, one direction in $l_2$ can significantly dominate others and slow down a minimization algorithm. As we are restricting iteration steps to the subspaces, we use shorthand notation $\|h\|_{x,\mathbf{S}} \overset{\text{def}}{=} \|h\|_{\nabla^2_{\mathbf{S}} f(x)}$.

## 2 ALGORITHM

### 2.1 Three Faces of the Algorithm

Our algorithm combines the best of three worlds (Table 2), and we can write it in three different ways.

**Theorem 1.** *If* $\nabla f(x^k)\in\text{Range}\left(\nabla^2 f(x^k)\right)$[3]*, then the following variants of the Newton method are equivalent:*

$$\textit{Regularize: } x^{k+1} = x^k + \mathbf{S}_k\operatorname*{argmin}_{h\in\mathbb{R}^d} T_{\mathbf{S}_k}\left(x^k,h\right), \tag{3}$$

$$\textit{Damping: } x^{k+1} = x^k - \alpha_k\mathbf{S}_k[\nabla^2_{\mathbf{S}_k} f(x^k)]^\dagger\nabla_{\mathbf{S}_k} f(x^k), \tag{4}$$

$$\textit{Sketching: } x^{k+1} = x^k - \alpha_k\mathbf{P}_{x^k}[\nabla^2 f(x^k)]^\dagger\nabla f(x^k), \tag{5}$$

*where* $\mathbf{P}_{x^k}$ *is a projection matrix onto* $\text{Range}\left(\mathbf{S}_k\right)$ *with respect to norm* $\|\cdot\|_{x_k}$ *(defined in eq. (9)),*

$$\begin{aligned}T_{\mathbf{S}}(x,h) &\overset{def}{=} f(x) + \langle\nabla f(x),\mathbf{S}h\rangle + \tfrac{1}{2}\|\mathbf{S}h\|^2_x + \tfrac{L_{alg}}{6}\|\mathbf{S}h\|^3_x\\ &= f(x) + \langle\nabla_{\mathbf{S}} f(x),h\rangle + \tfrac{1}{2}\|h\|^2_{x,\mathbf{S}} + \tfrac{L_{alg}}{6}\|h\|^3_{x,\mathbf{S}},\end{aligned} \tag{6}$$

$$\alpha_k \overset{def}{=} \frac{2}{1+\sqrt{1+2L_{alg}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}}. \tag{7}$$

[3] $\text{Range}\left(\mathcal{A}\right)$ denotes column space of the matrix $\mathcal{A}$.

Table 2: Three approaches for second-order global minimization. We denote $x^k \in \mathbb{R}^d$ iterates, $\mathbf{S}_k \sim \mathcal{D}$ distribution of sketches of rank $\tau \ll d$, $\alpha_k$ stepsizes. We report complexities for matrix inversions implemented naively.

| **Orthogonal lines of work** | **Sketch-and-project** (Gower and Richtárik, 2015) (various update rules) | **Damped Newton methods** (Nesterov and Nemirovski, 1994), (Karimireddy et al., 2018) | **Globally Regularized Newton** (Nesterov and Polyak, 2006), (Polyak, 2009), (Mishchenko, 2021), (Doikov and Nesterov, 2023) |
|---|---|---|---|
| **Update:** $x^{k+1} - x^k =$ | $= \alpha_k \mathbf{P}_{x^k}\left(\text{update}(x^k)\right)$ | $= \alpha_k [\nabla^2 f(x^k)]^\dagger \nabla f(x^k)$ | $= \text{argmin}_{h \in \mathbb{R}^d} T(x^k, h)$, for $T(x,h) \overset{\text{def}}{=} \langle \nabla f(x), h\rangle + \frac{1}{2}\|h\|_x^2 + \frac{L_2}{6}\|h\|_2^3$ |
| **Characteristics** | + cheap, low-rank updates<br>+ global linear convergence<br>– local quadratic rate unachievable | + affine-invariant geometry<br>– iteration cost $\mathcal{O}\left(d^3\right)$<br>Constant $\alpha_k$:<br>+ global linear convergence<br>Increasing $\alpha_k \nearrow 1$:<br>+ local quadratic rate | + global convex rate $\mathcal{O}\left(k^{-2}\right)$<br>+ local quadratic rate<br>– implicit updates[1]<br>– iteration cost $\mathcal{O}\left(d^3 \log \frac{1}{\varepsilon}\right)$[1] |

| **Combinations + retained benefits** | **Sketch-and-project** | **Damped Newton methods** | **Globally Regularized Newton** |
|---|---|---|---|
| RSN [9], Algorithm 3 | ✓<br>+ iter. cost $\mathcal{O}\left(d\tau^2\right)$<br>+ iter. cost $\mathcal{O}\left(1\right)$ if $\tau = 1$ | ✓<br>+ global linear rate | ✗ |
| SSCN [12], Algorithm 5 | ✓<br>+ iter. cost $\mathcal{O}\left(d\tau^2 + \tau^3 \log \frac{1}{\varepsilon}\right)$<br>+ iter. cost $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$ if $\tau = 1$<br>+ local linear rate $\mathcal{O}\left(\frac{d}{\tau} \log \frac{1}{\varepsilon}\right)$ | ✗ | ✓<br>+ global convex rate $\mathcal{O}\left(k^{-2}\right)$ |
| AICN [13], Algorithm 4 | ✗ | ✓<br>+ affine-invariant geometry<br>– no proof of global linear rate[3] | ✓<br>+ global convex rate $\mathcal{O}\left(k^{-2}\right)$<br>+ local quadratic rate<br>+ iteration cost $\mathcal{O}\left(d^3\right)$<br>+ simple, explicit updates |
| **SGN (this work)** Algorithm 1 | ✓<br>+ iter. cost $\mathcal{O}\left(d\tau^2\right)$<br>+ iter. cost $\mathcal{O}\left(1\right)$ if $\tau = 1$<br>+ local lin. rate $\mathcal{O}\left(\frac{d}{\tau} \log \frac{1}{\varepsilon}\right)$[2] | ✓<br>+ affine-invariant geometry<br>+ global linear rate | ✓<br>+ global convex rate $\mathcal{O}\left(k^{-2}\right)$<br>+ simple, explicit updates |

| **Three views of SGN** | **Sketch-and-project of damped Newton method** | **Damped Newton method in sketched subspaces** | **Affine-Invariant Newton in sketched subspaces** |
|---|---|---|---|
| **Update** $x^{k+1} - x^k =$ | $= \alpha_k \mathbf{P}_{x^k} [\nabla^2 f(x^k)]^\dagger \nabla f(x^k)$ | $= \alpha_k \mathbf{S}_k [\nabla_{\mathbf{S}_k} f(x^k)]^\dagger \nabla_{\mathbf{S}_k} f(x^k)$ | $= \mathbf{S}_k \, \text{argmin}_{h \in \mathbb{R}^d} T_{\mathbf{S}_k}(x^k, h)$, for $T_{\mathbf{S}}(x,h) \overset{\text{def}}{=} \langle \nabla f(x), \mathbf{S}h\rangle + +\frac{1}{2}\|\mathbf{S}h\|_x^2 + \frac{L_{\mathbf{S}}}{6}\|\mathbf{S}h\|_x^3$ |

[1] (Polyak, 2009; Mishchenko, 2021; Doikov and Nesterov, 2023) present algorithms with exact updates and cheaper per-iteration cost, but with certain tradeoffs. In particular, Polyak (2009) has slow $\mathcal{O}(k^{-1/4})$ global rate in convex regime. Mishchenko (2021) and Doikov and Nesterov (2023) have superlinear, but not quadratic local convergence.

[2] The rate is independent of the problem conditioning and faster than any first-order method.

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

*We call this algorithm Sketchy Global Newton, SGN, it is formalized as Algorithm 1.*

Note that $\alpha_k \in (0, 1]$ and $\alpha_k \to 1$ as the model $x^k$ converges to the solution. SGN enjoys a simpler convergence analysis compared to most of regularized Newton methods, as it allows an easy transition between gradient norms and model differences with the identity

$$\left\|x^{k+1} - x^k\right\|_{x^k} = \alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k, \mathbf{S}_k}. \tag{8}$$

## 2.2 Geometry of Sketches

We use a projection matrix on subspaces $\mathbf{S}$ with respect to local norms $\|\cdot\|_x$. Denote

$$\mathbf{P}_x \overset{\text{def}}{=} \mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x). \tag{9}$$

**Lemma 1.** *(Gower et al., 2020) Matrix $\mathbf{P}_x$ is a projection onto* Range $(\mathbf{S})$ *with respect to $\|\cdot\|_x$.*

We aim SGN to preserve Newton's direction in expectation. Form (5) shows that this holds as long as $\mathbf{S} \sim \mathcal{D}$ is such that $\mathbf{P}_x$ preserves direction in expectation.

**Assumption 1.** *Distribution $\mathcal{D}$ is chosen so that there*

*exists $\tau > 0$, such that*

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

**Lemma 2.** *Assumption 1 implies $\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\tau(\mathbf{S})\right] = \tau$.*

Assumption 1 is formulated in the local norms, so it might seem restrictive. To remediate that, in the experiment section, we demonstrate that in practice it can be omitted altogether. Moreover, in Section 5.3 we argue that such distribution can be constructed from simpler sketch distributions.

Projection matrix $\mathbf{P}_x$ from (9) has contractive properties, as stated by the following lemma.

**Lemma 3.** *Projection matrix $\mathbf{P}_x$ satisfies for any $g, h \in \mathbb{R}^d, g \in \text{Range}\left(\nabla^2 f(x)\right)$ inequalities*

$$\|\mathbf{P}_x h\|_x^2 \leq \|\mathbf{P}_x h\|_x^2 + \|(\mathbf{I}-\mathbf{P}_x)h\|_x^2 = \|h\|_x^2, \tag{11}$$

$$\mathbb{E}\left[\|\mathbf{P}_x h\|_x^2\right] = h^\top \nabla^2 f(x)\mathbb{E}\left[\mathbf{P}_x\right] h \overset{As.1}{=} \frac{\tau}{d}\|h\|_x^2, \tag{12}$$

$$\mathbb{E}\left[\left\|\mathbf{P}_x^\top g\right\|_x^{*2}\right] = g^\top \mathbb{E}\left[\mathbf{P}_x\right]\left[\nabla^2 f(x)\right]^\dagger g \overset{As.1}{=} \frac{\tau}{d}\|g\|_x^{*2}, \tag{13}$$

$$\mathbb{E}\left[\|\mathbf{P}_x h\|_x^3\right] \overset{(11)}{\leq} \mathbb{E}\left[\|h\|_x \cdot \|\mathbf{P}_x h\|_x^2\right] \overset{(12)}{=} \frac{\tau}{d}\|h\|_x^3. \tag{14}$$

### 2.3 Affine-Invariant Assumptions

To leverage the affine-invariance of the norms, we use an affine-invariant version of second-order smoothness called *semi-strong self-concordance* (Hanzely et al., 2022).

**Definition 1.** *A twice differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$ is called $L_{semi}-$semi-strongly self-concordant if for $L_{semi} < \infty$ holds*

$$\left\|\nabla^2 f(y) - \nabla^2 f(x)\right\|_{op} \leq L_{semi}\|y-x\|_x, \quad \forall y, x \in \mathbb{R}^d,$$

*where operator norm is, for given $x \in \mathbb{R}^d$, defined for any matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ as*

$$\|\mathbf{H}\|_{op} \overset{def}{=} \sup_{v\in\mathbb{R}^d} \frac{\|\mathbf{H}v\|_x^*}{\|v\|_x}.$$

Semi-strong self-concordance differs from the standard second-order smoothness only in the norm in which the distance is measured. It follows from the Lipschitz smoothness and the strong convexity.

In Section 3.3 we relax the Definition 1 to the *self-concordance* (Nesterov and Nemirovski, 1994), defined below. To allow tighter constants, we define it separately for each sketched subspace.

**Definition 2.** *A three times differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$, is called $L_{\mathbf{S}}$–self-concordant in range of $\mathbf{S}$ if*

$$\left|\nabla^3 f(x)[\mathbf{S}h]^3\right|, \leq L_{\mathbf{S}}\|\mathbf{S}h\|_x^3 \quad \forall x \in \mathbb{R}^d, h \in \mathbb{R}^{\tau(\mathbf{S})}\setminus\{0\},$$

*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 $h \in \mathbb{R}^d$.*

We refer the reader for the more detailed comparison of smoothness assumptions to Appendix D.

### 2.4 One-Step Decrease

Self-concordance implies a standard descent lemma.

**Lemma 4.** *SGN step (4) decreases loss of a $L_{\mathbf{S}}$-self-concordant function $f : \mathbb{R}^d \to \mathbb{R}$ as*

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2}\min\left\{\left(L_{alg}g_k\right)^{-\frac{1}{2}}, \frac{1}{2}\right\} g_k^2, \tag{15}$$

*where $g_k = \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^*$.*

Therefore, all iterates of SGN stay in the initial level set $\mathcal{Q}(x_0) \overset{\text{def}}{=} \left\{x \in \mathbb{R}^d : f(x) \leq f(x^0)\right\}$. However, the decrease of (15) does not lead to the fast global $\mathcal{O}\left(k^{-2}\right)$ rate. Instead, we are going to analyze SGN as a regularized Newton method. We show $T_{\mathbf{S}}(x,h)$ upper bounds the function $f$ and quantify the one-step decrease.

**Proposition 1** (Hanzely et al. (2022), Lemma 2)**.** *For $L_{semi}$–semi-strong self-concordant $f : \mathbb{R}^d \to \mathbb{R}$, 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 $f(x_+) \leq T_{\mathbf{S}}(x,h)$ and*

$$\left|f(x_+) - f(x) - \langle\nabla f(x), \mathbf{S}h\rangle - \tfrac{1}{2}\|\mathbf{S}h\|_x^2\right| \leq \tfrac{L_{semi}}{6}\|\mathbf{S}h\|_x^3 \tag{16}$$

**Lemma 5.** *Fix any $y \in \mathbb{R}^d$. Let the function $f : \mathbb{R}^d \to \mathbb{R}$ 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 the one-step decrease*

$$\mathbb{E}\left[f(x^{k+1})|x^k\right] \leq \left(1-\frac{\tau}{d}\right) f(x^k) + \frac{\tau}{d}f(y) + \frac{\tau}{d}\frac{L_{alg}+L_{semi}}{6}\left\|y-x^k\right\|_{x^k}^3. \tag{17}$$

With Lemma 5, we are one step before the main converge result. All that is left is to choose $y$ as a linear combination of $x^k$ and $x^*$ and to bound distance between $\left\|x^k - x^*\right\|_{x^k}$.

**Remark.** *Proposition 1 also implies loss decrease. For $h^* \overset{def}{=} \text{argmin}_{h\in\mathbb{R}^{\tau(\mathbf{S})}} T_{\mathbf{S}}(x,h)$ and $x_+ \overset{def}{=} x + \mathbf{S}h$ holds*

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

## 3 CONVERGENCE RESULTS

Finally, we are ready to present the main convergence results: the fast global $\mathcal{O}\left(k^{-2}\right)$ rate, the fast local conditioning-independent linear rate, and the global linear convergence rate to the neighborhood of the solution.

Table 3: Globally convergent Newton-like methods for $d$-dimensional, smooth, strongly convex functions. We highlight the best-known rates in blue.

| Algorithm | Stepsize range | Affine-invariant algorithm? | Iteration cost[0] | Linear rate[1] | Global convex rate | 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 | ✗ | $\mathcal{O}\left(d^3 \log \frac{1}{\varepsilon}\right)$[3] | ✗ | $\mathcal{O}\left(k^{-2}\right)$ | (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, 2023) |
| Exact Newton Descent (Alg. 2) | $\frac{1}{\hat{L}}$[4] | ✓ | $\mathcal{O}\left(d^3\right)$ | global | ✗ | (Karimireddy et al., 2018) |
| RSN (Algorithm 3) | $\frac{1}{\hat{L}}$[4] | ✓ | $\mathcal{O}(d\tau^2)$, $\mathcal{O}(1)$ if $\tau=1$ | global | $\mathcal{O}\left(k^{-1}\right)$ | (Gower et al., 2019) |
| SSCN Algorithm 5 | 1 | ✗ | $\mathcal{O}(d\tau^2+\tau^3 \log \frac{1}{\varepsilon})$,[3] $\mathcal{O}(\log \frac{1}{\varepsilon})$ if $\tau=1$ | local | $\mathcal{O}\left(k^{-1}\right)$ | (Hanzely et al., 2020) |
| **SGN** (Algorithm 1) | $(0,1]$ | ✓ | $\mathcal{O}(d\tau^2)$, $\mathcal{O}(1)$ if $\tau=1$ | global[5] +local | $\mathcal{O}\left(k^{-2}\right)$ | **This work** |

[0] Constant $\tau$ rank of sketches matrices $\mathbf{S}_k$, $\tau \ll d$. We report the rate of implementation using matrix inverses.
[1] Terms "local" and "global" denote whether algorithm has local/global linear rate (under possibly stronger assumptions).
[3] Cubic Newton and SSCN solve an implicit problem in each iteration, which naively implemented, requires $\times \log \frac{1}{\varepsilon}$ matrix inverses approximate sufficiently (Hanzely et al., 2022). For $\tau \log \frac{1}{\varepsilon} \geq d$ (larger $\tau$ or high precision $\varepsilon$), this becomes the bottleneck of SSCN.
[4] Under $\hat{L}$-relative smoothness (Def. 3). Karimireddy et al. (2018) utilizes similar $c$-stability assumption instead.
[5] Separate results for local convergence (Th. 3) and global convergence to the corresponding neighborhood (Th. 4).

### 3.1 Global Convex $\mathcal{O}\left(k^{-2}\right)$ Convergence

Global convergence rate depends on the diameter of the initial level $\mathcal{Q}(x_0)$ defined as

$$R \overset{\text{def}}{=} \sup_{x,y \in \mathcal{Q}(x_0)} \|x-y\|_x.$$

**Theorem 2.** *Let $L_{semi}$–semi-strongly concordant function $f:\mathbb{R}^d \to \mathbb{R}$ be with a finite diameter of initial level set $\mathcal{Q}(x_0)$, $R<\infty$, and sketching matrices satisfy Assumption 1. SGN has global $\mathcal{O}(k^{-2})$ convergence,*

$$\mathbb{E}\left[f(x^k)-f^*\right] \leq \frac{4d^3(f(x^0)-f^*)}{\tau^3 k^3} + \frac{9(\max L_{alg}+L_{semi})d^2R^3}{2\tau^2 k^2}.$$

### 3.2 Fast Local Linear Convergence

Near the solution, SGN achieves linear convergence independent of problem conditioning. Such result is tight since, as a sketch-and-project method, SGN cannot attain a local superlinear rate (see Section 5.1).

**Theorem 3.** *Let function $f:\mathbb{R}^d \to \mathbb{R}$ be $L_{\mathbf{S}}$–self-concordant in subspaces $\mathbf{S} \sim \mathcal{D}$ and expected projection matrix unbiased (Assumption 1). Iterates of SGN $x^0,\ldots,x^k$ such that[4] $\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \leq \frac{1}{L_{semi}}$ and $\nabla f(x^k) \in \text{Range}\left(\nabla^2 f(x^k)\right)$ have local linear rate*

$$\mathbb{E}\left[f(x^k)-f^*\right] \leq \left(1-\frac{\tau}{bd}\right)^k (f(x^0)-f^*), \qquad (18)$$

*for $b \overset{def}{=} \max\left\{\sqrt{\frac{L_{alg}}{L_{semi}}}, 2\right\}$, and the local complexity of SGN is independent on the problem conditioning, $\mathcal{O}\left(\sqrt{\frac{L_{alg}}{L_{semi}}}\frac{d}{\tau}\log\frac{1}{\varepsilon}\right)$ and $\mathcal{O}\left(\frac{d}{\tau}\log\frac{1}{\varepsilon}\right)$ for $L_{alg}=L_{semi}$.*

[4] This can be relaxed by replacing $L_{\text{semi}}$ with $L_{\mathbf{S}_k}$.

### 3.3 Global Linear Convergence

Our last convergence result is a global linear rate under relative smoothness and relative convexity.

**Definition 3.** *(Gower et al., 2019) We call relative convexity and relative smoothness in subspace* $\mathbf{S}$ *constants* $\hat{\mu}, \hat{L}_{\mathbf{S}} > 0$, *s.t. for all* $x, y \in \mathbb{R}^d, h \in \mathbb{R}^{\tau(\mathbf{S})}$ *and* $y_{\mathbf{S}} = x + \mathbf{S}h$ *hold*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\hat{\mu}}{2}\|y - x\|_x^2, \tag{19}$$

$$f(y_{\mathbf{S}}) \leq f(x) + \langle \nabla_{\mathbf{S}} f(x), y_{\mathbf{S}} - x \rangle + \frac{\hat{L}_{\mathbf{S}}}{2}\|y_{\mathbf{S}} - x\|_{x,\mathbf{S}}^2. \tag{20}$$

Gower et al. (2019) shows that updates $x_+ = y_{\mathbf{S}}$, where $y_{\mathbf{S}}$ is a minimizer of RHS of (20) can be written as Newton method with stepsize $\frac{1}{\hat{L}}$ and have global linear convergence. Conversely, our stepsize $\alpha_k$ varies, (7), so this result is not directly applicable to us. Surprisingly, it turns out that a careful choice of $L_{\text{alg}}$ can guarantee global linear convergence. Observe following:

- We can write SGN model (6) similarly to the relative smoothness (20),

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

  If $\left(1 + \frac{L_{\text{alg}}}{3}\|h\|_{x,\mathbf{S}}\right) \geq \hat{L}_{\mathbf{S}}$, then SGN model upper bounds the right-hand side of (20) and therefore, function $f$ as well. Consequently, we can obtain rates similar to Gower et al. (2019).
- To guarantee $\left(1 + \frac{L_{\text{alg}}}{3}\|h\|_{x,\mathbf{S}}\right) \geq \hat{L}_{\mathbf{S}}$, we can express $L_{\text{alg}}$ using $\|h\|_{x,\mathbf{S}} = \alpha\|g\|_{x,\mathbf{S}}^*$ as:

$$\begin{aligned} & 1 + \frac{L_{\text{alg}}}{3}\|h\|_{x,\mathbf{S}} \geq \hat{L}_{\mathbf{S}} \\ \Leftrightarrow\ & L_{\text{alg}} \geq \frac{3(\hat{L}_{\mathbf{S}} - 1)}{\alpha\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*} \\ \Leftrightarrow\ & 1 \geq \frac{3(\hat{L}_{\mathbf{S}} - 1)}{-1 + \sqrt{1 + 2L_{\text{alg}}\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*}} \\ \Leftrightarrow\ & L_{\text{alg}} \geq \frac{3}{2}\frac{(\hat{L}_{\mathbf{S}} - 1)(3\hat{L}_{\mathbf{S}} - 1)}{\|\nabla_{\mathbf{S}} f(x)\|_{x,\mathbf{S}}^*}. \end{aligned}$$

- We have already shown the fast local convergence of SGN (Theorem 3). Now we need to obtain linear rate for just points $x^k$ beyond that neighborhood of convergence, $\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^* \geq \frac{1}{L_{\mathbf{S}_k}}$.

  For those points $x^k$ we can guarantee it with the choice $L_{\text{alg}} \geq \sup_{\mathbf{S}} \frac{9}{2} L_{\mathbf{S}} \hat{L}_{\mathbf{S}}^2$. Such $L_{\text{alg}}$ also bounds the stepsize far from the solution as $\alpha_k \hat{L}_{\mathbf{S}_k} \leq \frac{2}{3}$ (see Lemma 10 in Appendix E.12).

We are almost ready to present the global linear convergence result. Finally, the rate depends on the conditioning of the expected projection matrix $\mathbf{P}_x$, defined as[5]

$$\rho(x) = \min_{g \in \mathbb{R}^d} \frac{g^\top \mathbb{E}\left[\alpha \mathbf{P}_x\right] [\nabla^2 f(x)]^\dagger g}{\|g\|_x^{*2}} \tag{21}$$

$$\rho \overset{\text{def}}{=} \min_{x \in \mathcal{Q}(x_0)} \rho(x). \tag{22}$$

**Theorem 4.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be* $\hat{L}_{\mathbf{S}}$*–relative smooth in subspaces* $\mathbf{S}$ *and* $\hat{\mu}$*–relative convex. Let sampling* $\mathbf{S} \sim \mathcal{D}$ *satisfy* $\text{Null}\left(\mathbf{S}^\top \nabla^2 f(x) \mathbf{S}\right) = \text{Null}\left(\mathbf{S}\right)$ *and* $\text{Range}\left(\nabla^2 f(x)\right) \subset \text{Range}\left(\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{S}\mathbf{S}^\top\right]\right)$. *Then* $0 < \rho \leq 1$. *Choose parameter* $L_{alg} = \sup_{\mathbf{S}\sim\mathcal{D}} \frac{9}{2} L_{\mathbf{S}} \hat{L}_{\mathbf{S}}^2$.

*While iterates* $x^0, \ldots, x^k$ *satisfy* $\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^* \geq \frac{1}{L_{\mathbf{S}_k}}$, *then*[6] *SGN has the decrease*

$$\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{4}{3}\rho\hat{\mu}\right)^k (f(x^0) - f^*),$$

*and global linear* $\mathcal{O}\left(\frac{1}{\rho\hat{\mu}} \log \frac{1}{\varepsilon}\right)$ *convergence.*

## 4 EXPERIMENTS

We support our theory by comparing SGN to SSCN. To match practical considerations of SSCN and for the sake of simplicity, we adjust SGN in unfavorable way:

1. We choose sketching matrices $\mathbf{S}$ to be unbiased in $l_2$ norms (instead of local hessian norms $\|\cdot\|_x$ from Assumption 1),
2. To disregard implementation specifics, we report iterations on the $x$–axis. Note that SSCN needs to use a subsolver (extra line-search) to solve the implicit step in each iteration. If naively implemented using matrix inverses, iterations of SSCN are $\times \log \frac{1}{\varepsilon}$ slower. We haven't reported time as this would naturally ask for hardware-optimized implementations, which was out of the scope of the paper.

Despite the simplicity of SGN and unfavorable adjustments, Figure 1 shows that SGN performs comparably to SSCN. Figure 2 presents a comparison of SGN, SSCN, CD, and Accelerated Coordinate Descent.

[5] We formulate the condition number $\rho(x)$ in local norms, but $l_2$ norms can be used as well.

[6] If opposite inequality holds in the current iterate, then Theorem 3 guarantees even faster convergence.

We can point out other properties of SGN based on experiments in the literature.

- **Rank of 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 (the Newton method with the fixed stepsize) is much faster than first-order Accelerated Coordinate Descent (ACD) for highly dense problems. For extremely sparse problems, ACD has competitive performance. As the stepsize of SGN increases as converging 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 ACD. As SGN is competitive to SSCN, similar results are expected to 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 analogical updates (in the subspaces), it indicates that SGN saves time over SSCN.

As a measure of convergence, we report shifted and rescaled functional values measure called *relative subpotimality*, which can be calculated at point $x$ as $\frac{f(x)-f^*}{f(x^0)-f^*}$.

# 5 DISCUSSION

## 5.1 Local Linear Convergence Limit

Similarly to AICN (Hanzely et al., 2022), we can show that one step decreases the gradient norm quadratically. In our case, the quadratic decrease is the sketched subspace.

**Lemma 6.** *For $L_{semi}$–semi-strong self-concordant function $f : \mathbb{R}^d \to \mathbb{R}$ and parameter choice $L_{alg} \geq L_{semi}$, one step of SGN has quadratic decrease in the subspace* Range $(\mathbf{S}_k)$,

$$\left\|\nabla_{\mathbf{S}_k} f(x^{k+1})\right\|^*_{x^k,\mathbf{S}_k} \leq L_{alg}\alpha_k^2 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*2}_{x^k,\mathbf{S}_k}. \quad (23)$$

Nevertheless, this is insufficient for superlinear local convergence; we can achieve a linear rate at best. We can illustrate this on an edge case where $f : \mathbb{R}^d \to \mathbb{R}$ is a quadratic function: self-concordance assumption holds with $L_{\mathbf{S}} = 0$ and as $\alpha_k \xrightarrow{L_{\mathbf{S}} \to 0} 1$, SGN stepsize becomes 1 and SGN simplifies to subspace Newton method. Unfortunately, the subspace Newton method has just linear local convergence (Gower et al., 2019; Hanzely et al., 2020).

## 5.2 Is Rate $\mathcal{O}\left(k^{-2}\right)$ Possible for SSCN?

Hanzely et al. (2020) proposed SSCN, the sketch-and-project version of the cubic Newton method. While it intuitively seems that directly combining these approaches could directly lead to the desired global rate of $\mathcal{O}(k^{-2})$, achieving such a rate demands an extremely careful choice of assumptions and distribution of sketch matrices. Unfortunately, SSCN has a slight mismatch, resulting in a slower rate of $\mathcal{O}(k^{-1})$. We can present a slight modification of the SSCN algorithm to showcase what modification is needed to achieve the desired global rate of $\mathcal{O}(k^{-2})$.

For functions $f$ that satisfy for $\forall \mathbf{S} \sim \mathcal{D}, \forall h \in \mathbb{R}^{\tau(\mathbf{S})}$

$$\left|f(x+\mathbf{S}h) - f(x) - \langle \nabla f(x), \mathbf{S}h\rangle - \tfrac{1}{2}\|\mathbf{S}h\|_2^2\right| \leq \tfrac{L'}{6}\|\mathbf{S}h\|_2^3, \quad (24)$$

the sequence of iterates is defined as $x^{k+1} =$

$$\operatorname*{argmin}_{h\in\mathbb{R}^d} \left\{ f(x^k) + \langle \nabla f(x), \mathbf{S}h\rangle + \tfrac{1}{2}\|\mathbf{S}h\|_2^2 + \tfrac{L'_{est}}{6}\|\mathbf{S}h\|_2^3 \right\}, \quad (25)$$

for constant $L'_{est} \geq L'$ and sketch $\mathbf{S}$ sampled from distribution $\mathcal{D}$, s.t.

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_2 \overset{\text{def}}{=} \mathbf{S}(\mathbf{S}^\top\mathbf{S})^\dagger\mathbf{S}^\top\right] = \frac{\tau}{d}\mathbf{I}, \quad (26)$$

implying $\mathbb{E}[\|\mathbf{P}_2 h\|_2^2] = \frac{\tau}{d}\|\mathbf{h}\|_2^2$, achieves global convex convergence rate $\mathcal{O}(k^{-2})$. However, while the left-hand side of (16) is the second-order Taylor expansion, the left-hand side of (24) is not. We currently do not know which class of functions $f$ satisfies this requirement.

In the case of the original SSCN, there is a discrepancy between usage of local norms in the left-hand side of (24) and $l_2$ norms in the update rule (25). This causes an extra quadratic term in (17), ultimately resulting in a slower convergence rate $\mathcal{O}(k^{-1})$.

## 5.3 Construction of the Sketch Distribution

Here we demonstrate that the distribution of sketching matrices satisfying Assumption 1 can be obtained from sketches with $l_2$–unbiased projection (which were used in (Hanzely et al., 2020)).

**Lemma 7** (Construction of sketch matrix $\mathbf{S}$)**.** *If we have a sketch matrix distribution $\tilde{\mathcal{D}}$ so that a projection on* Range $(\mathbf{M})$, $\mathbf{M} \sim \tilde{\mathcal{D}}$ *is unbiased in $l_2$ norms,*

$$\mathbb{E}_{\mathbf{M}\sim\tilde{\mathcal{D}}}\left[\mathbf{M}\left(\mathbf{M}^\top\mathbf{M}\right)^\dagger\mathbf{M}^\top\right] = \frac{\tau}{d}\mathbf{I}, \quad (27)$$

*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*

$$\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_x\right] = \frac{\tau}{d}\mathbf{I}. \quad (28)$$

A more practical way to sample from the distribution sketch distribution is subject to future research.

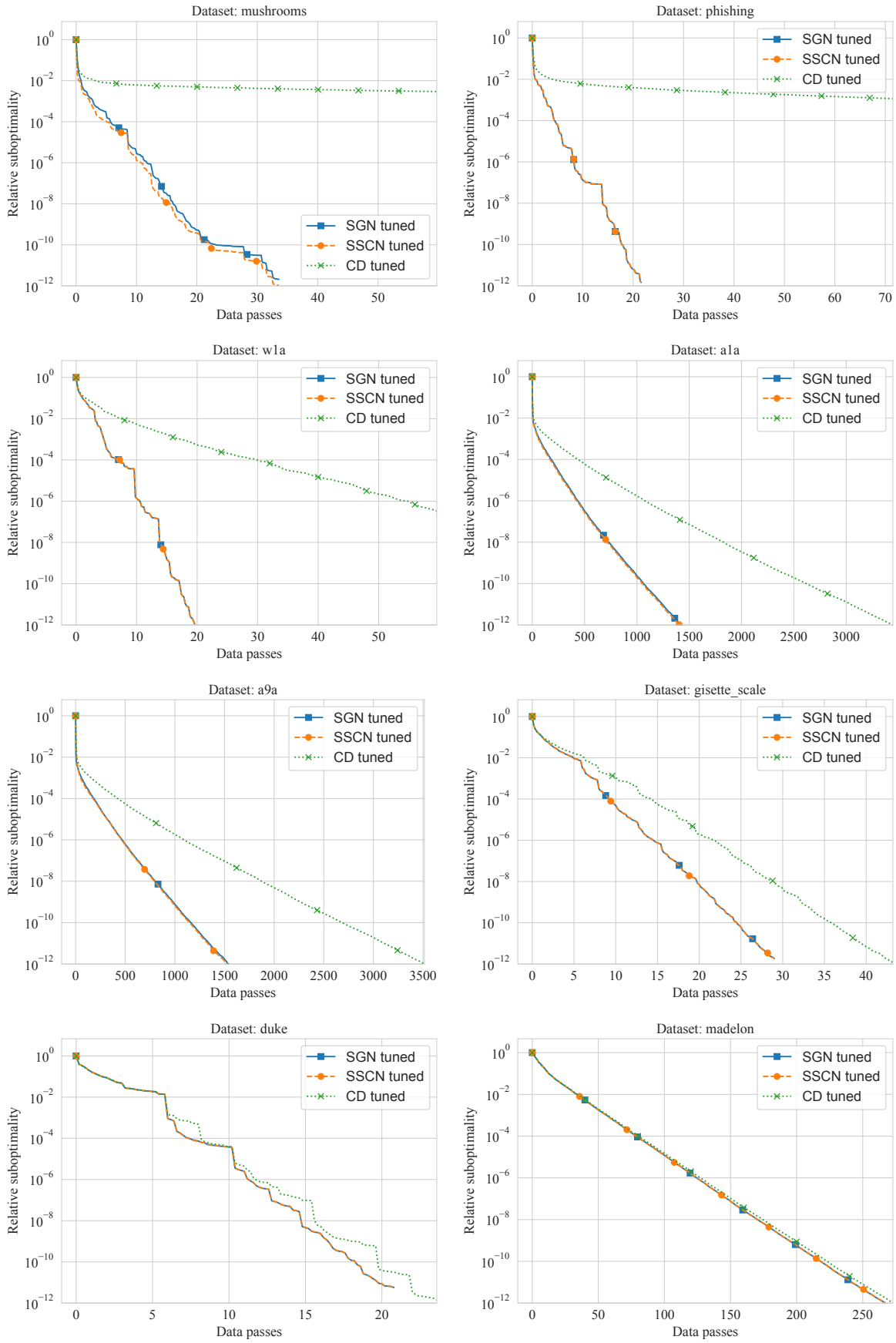


Figure 1: Comparison of SSCN, SGN and CD on the logistic regression loss on LIBSVM datasets for sketch matrices $\mathbf{S}$ of rank one. We fine-tune all algorithms for their smoothness parameters.

## Acknowledgements

The work of Slavomír Hanzely was supported by the King Abdullah University of Science and Technology (KAUST) through the baseline fund BAS/1/1677-01-01 of Professor Peter Richtárik.

### References

Jérôme Bolte and Edouard Pauwels. Curiosities and counterexamples in smooth convex optimization. *Mathematical Programming*, 195(1-2):553–603, 2022.

Bruce Christianson. Automatic Hessians by reverse accumulation. *IMA Journal of Numerical Analysis*, 12(2):135–150, 1992.

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

Nikita Doikov and Yurii Nesterov. Local convergence of tensor methods. *Mathematical Programming*, 193: 315–336, 2022.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical Programming*, pages 1–25, 2023.

Nikita Doikov and Peter Richtárik. Randomized block cubic Newton method. In Jennifer Dy and Andreas Krause, editors, *The 35th International Conference on Machine Learning*, 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 Gower and Margarida Mello. A new framework for the computation of Hessians. *Optimization Methods and Software*, 27(2):251–273, 2012.

Robert Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Robert 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 Systems32*, pages 616–625. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/8351-rsn-randomized-subspace-newton.pdf`.

Robert Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

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 $\mathcal{O}(k^{-2})$ and local quadratic convergence rate. *Advances in Neural Information Processing Systems*, 35: 25320–25334, 2022.

Florian Jarre and Philippe 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 Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6): 89–185, 1948.

Sai Karimireddy, Sebastian Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint: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 Mascarenhas. On the divergence of line search methods. *Computational & Applied Mathematics*, 26 (1):129–169, 2007.

Konstantin Mishchenko. Regularized Newton method with global $\mathcal{O}(1/k^2)$ convergence. *arXiv preprint:2112.02089*, 2021.

Jorge 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 Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

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

Roman 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*, 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.

Anton Rodomanov and Yurii Nesterov. Greedy quasi-Newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.

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 Ypma. Historical development of the Newton–Raphson method. *SIAM Review*, 37(4):531–551, 1995.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Appendix

## A TABLE OF FREQUENTLY USED NOTATION

Table 4: Summary of frequently used notation

| | |
|---|---|
| **General** | |
| $\mathbf{A}^\dagger$ | Moore-Penrose pseudoinverse of $\mathbf{A}$ |
| $\|\cdot\|_{op}$ | Operator norm |
| $\|\cdot\|_x$ | Local norm at $x$ |
| $\|\cdot\|_x^*$ | Local dual norm at $x$ |
| $x, x_+, x^k \in \mathbb{R}^d$ | Iterates |
| $y \in \mathbb{R}^d$ | Virtual iterate (for analysis only) |
| $h, h' \in \mathbb{R}^d$ | Difference between consecutive iterates |
| $\alpha_k$ | SGN Stepsize |
| **Function specific** | |
| $d$ | Dimension of problem |
| $f : \mathbb{R}^d \to \mathbb{R}$ | Loss function |
| $T_{\mathbf{S}}(\cdot, x)$ | Upperbound on $f$ based on gradient and Hessian in $x$ |
| $x^*,\ f^*$ | Optimal model and optimal function value |
| $\mathcal{Q}(x_0)$ | Set of models with a functional value less than $x^0$ |
| $R, D, D_2$ | Diameter of $\mathcal{Q}(x_0)$ |
| $L_{\text{sc}}, L_{\text{semi}}$ | Self-concordance and semi-strong self-concordance constants |
| $L_{\text{alg}}$ | Smoothness estimate, affects stepsize of SGN |
| $\hat{L}, \hat{\mu}$ | Relative smoothness and relative convexity constants |
| **Sketching** | |
| $\nabla_{\mathbf{S}} f, \nabla^2_{\mathbf{S}} f, \|h\|_{x,\mathbf{S}}$ | Gradient, Hessian, local norm in range $\mathbf{S}$, resp. |
| $\mathbf{S} \in \mathbb{R}^{d \times \tau(\mathbf{S})}$ | Randomized sketching matrix |
| $\tau(\mathbf{S})$ | Dimension of randomized sketching matrix |
| $\tau$ | Fixed dimension constraint on $\mathbf{S}$ |
| $L_{\mathbf{S}}$ | Self-concordance constant in range of $\mathbf{S}$ |
| $\mathbf{P}_x$ | Projection matrix on subspace $\mathbf{S}$ w.r.t. local norm at $x$ |
| $\rho(x)$ | Condition numbers of expected scaled projection matrix $\mathbb{E}[\alpha \mathbf{P}_x]$ |
| $\rho$ | Lower bound on condition numbers $\rho(x)$ |

## B TECHNICAL DETAILS OF EXPERIMENTS

We use a 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 a 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 a random seed, experiments are fully reproducible.

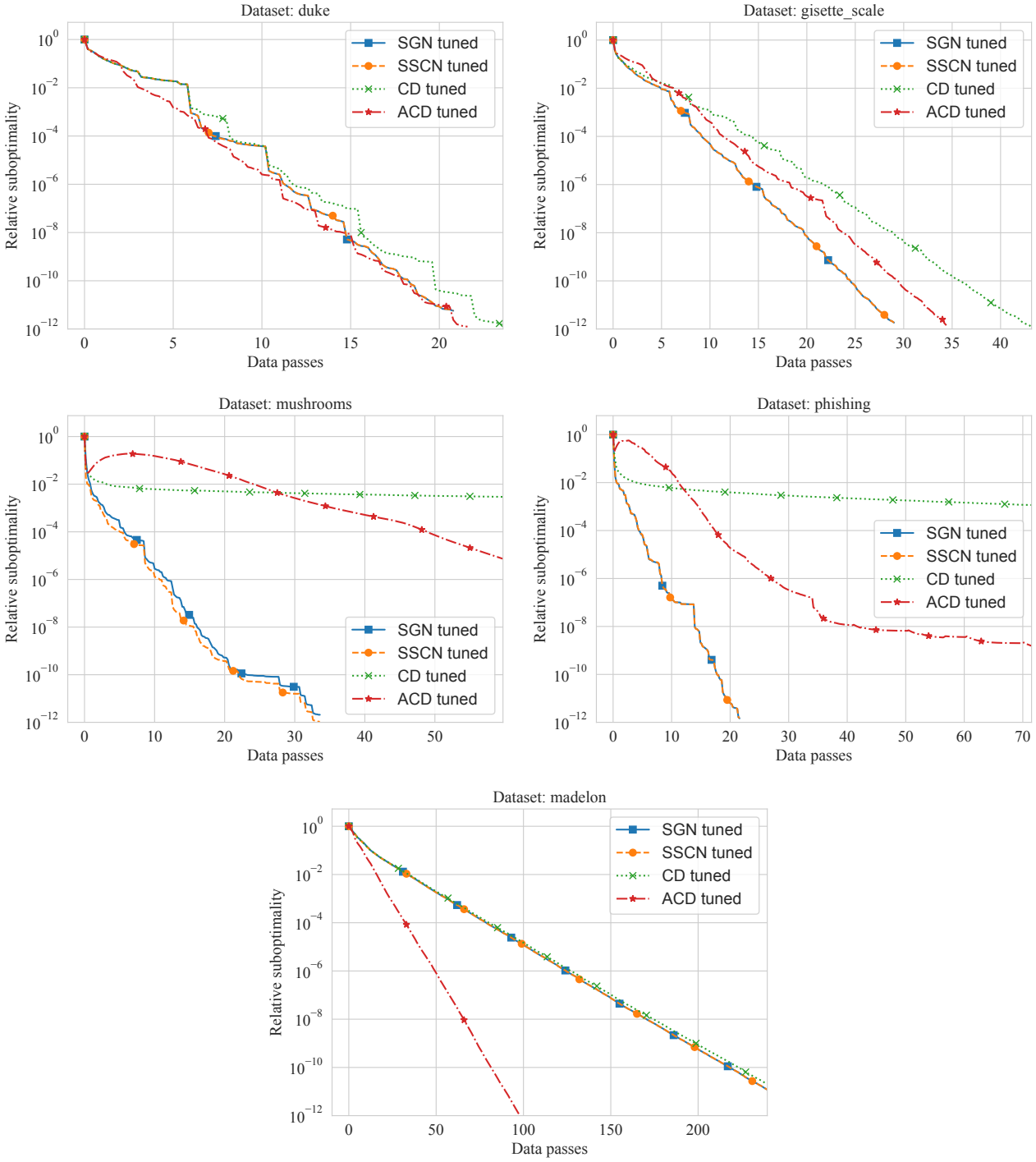


Figure 2: Comparison of SSCN, SGN, CD and ACD on logistic regression on LIBSVM datasets for sketch matrices $\mathbf{S}$ of rank one. We fine-tune all algorithms for smoothness parameters.

## C ALGORITHM COMPARISON

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

**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** RSN: 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 $\mathbf{S}_k \sim \mathcal{D}$

$x^{k+1} = x^k - \frac{1}{L} \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x^k)$

**end for**

**Algorithm 4** AICN: Affine-Invariant Cubic Newton (Hanzely et al., 2022)

**Requires:** Initial point $x^0 \in \mathbb{R}^d$, semi-strong self-concordance upper bound $L_{\text{alg}} \geq L_{\text{semi}} > 0$

**for** $k = 0, 1, 2 \dots$ **do**

$\alpha_k = \frac{2}{1+\sqrt{1+2L_{\text{alg}}\|\nabla f(x^k)\|^*_{x^k}}}$

$x^{k+1} = x^k - \alpha_k \left[\nabla^2 f(x^k)\right]^\dagger \nabla f(x^k)$ [a]

**end for**

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

**Algorithm 5** SSCN: Stochastic Subspace Cubic Newton (Hanzely et al., 2020)

**Requires:** Initial point $x^0 \in \mathbb{R}^d$, distribution of random matrices $\mathcal{D}$, semi-strong self-concordance upper bound $L_{\text{alg}} \geq L_{\mathbf{S}} > 0$

**for** $k = 0, 1, 2 \dots$ **do**

Sample $\mathbf{S}_k \sim \mathcal{D}$

$\alpha_k = \frac{2}{1+\sqrt{1+2L_{\text{alg}}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k, \mathbf{S}_k}}}$

$x^{k+1} = x^k - \alpha_k \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x^k)$ [a]

**end for**

[a] Equivalently, $x^{k+1} = x^k - \mathbf{S}_k \operatorname{argmin}_{h \in \mathbb{R}^d} \hat{T}_{\mathbf{S}_k}(x^k, h)$, for $\hat{T}_{\mathbf{S}}(x, h) = \langle \nabla f(x), \mathbf{S}h\rangle + \frac{1}{2}\|\mathbf{S}h\|^2_x + \frac{L_{\mathbf{S}}}{6}\|\mathbf{S}h\|^3$.

Figure 3: Pseudocodes of algorithms related to SGN. We highlight the stepsizes of the Newton method in blue, subspace sketching in green, and regularized Newton step in brown. We present AICN in the simplified form.

## D SELF-CONCORDANCE OVERVIEW

Self-concordance (Nesterov and Nemirovski, 1994) is a variant of the smoothness expressed in local norms.

**Definition 4.** *Convex function $f$ with continuous first, second, and third derivatives is called self-concordant if*

$$|D^3 f(x)[h]^3| \leq L_{sc}\|h\|^3_x, \quad \forall x, h \in \mathbb{R}^d, \tag{29}$$

*where for any integer $p \geq 1$, by $D^p f(x)[h]^p \overset{def}{=} D^p f(x)[h, \dots, h]$ we denote the $p$–th order directional derivative[7] of $f$ at $x \in \mathbb{R}^d$ along direction $h \in \mathbb{R}^d$.*

This assumption corresponds to a big class of optimization methods called interior-point methods (Nesterov and Nemirovski, 1994), it implies the uniqueness of the solution of the lower bounded function (Nesterov et al., 2018, Theorem 5.1.16).

Similarly to Lipschitz smoothness, self-concordance implies an upper bound on the function value; however, in local norms.

$$f(y) - f(x) \leq \langle \nabla f(x), y - x\rangle + \frac{1}{2}\|y - x\|^2_x + \frac{L_{\text{sc}}}{6}\|y - x\|^3_x, \qquad \forall x, y \in \mathbb{R}^d. \tag{30}$$

Note that this definition matches the definition of self-concordance in sketched subspaces (Definition 2) with $\mathbf{S} = \mathbf{I}$, and consequently $L_{\mathbf{S}} \leq L_{\text{sc}}$.

[7] For example, $D^1 f(x)[h] = \langle \nabla f(x), h\rangle$ and $D^2 f(x)[h]^2 = \langle \nabla^2 f(x)h, h\rangle$.

Going beyond self-concordance, Rodomanov and Nesterov (2021) introduced a stronger version of the self-concordance assumption.

**Definition 5.** *Twice differentiable convex function, $f \in C^2$, is called strongly self-concordant if*

$$\nabla^2 f(y) - \nabla^2 f(x) \preceq L_{str}\|y-x\|_z \nabla^2 f(w), \quad \forall y, x, z, w \in \mathbb{R}^d. \tag{31}$$

In this paper, we are working with the class of semi-strong self-concordant functions (Hanzely et al., 2022)

$$\left\|\nabla^2 f(y) - \nabla^2 f(x)\right\|_{op} \leq L_{\text{semi}}\|y-x\|_x, \quad \forall y, x \in \mathbb{R}^d, \tag{32}$$

which is analogous to standard second-order smoothness

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq L_2\|x-y\|. \tag{33}$$

All of the mentioned self-concordance variants are affine-invariant and their respective classes satisfy (Hanzely et al., 2022)

$$\textit{strong self-concordance} \subseteq \textit{semi-strong self-concordance} \subseteq \textit{self-concordance}.$$

Also, for a fixed strongly self-concordant function $f$ and smallest such $L_{\text{sc}}, L_{\text{semi}}, L_{\text{str}}$ holds $L_{\text{sc}} \leq L_{\text{semi}} \leq L_{\text{str}}$ (Hanzely et al., 2022).

All notions of self-concordance are closely related to the standard convexity and smoothness; Rodomanov and Nesterov (2021) shows that strong self-concordance follows from function $L_2$–Lipschitz continuous Hessian and strong convexity.

**Proposition 2.** *(Rodomanov and Nesterov, 2021, Example 4.1) Let $\mathbf{H} : \mathbb{R}^d \to \mathbb{R}^d$ be a self-adjoint positive definite operator. Suppose there exist $\mu > 0$ and $L_2 \geq 0$ such that the function $f$ is $\mu$–strongly convex and its Hessian is $L_2$–Lipschitz continuous* (33) *with respect to the norm $\|\cdot\|_{\mathbf{H}}$. Then $f$ is strongly self-concordant with constant $L_{str} = \frac{L_2}{\mu^{3/2}}$.*

# E MISSING PROOFS

## E.1 Proof of Theorem 1

*Proof.* Because $\nabla f(x^k) \in \text{Range}\left(\nabla^2 f(x^k)\right)$, it holds $\nabla^2 f(x^k)[\nabla^2 f(x^k)]^\dagger \nabla f(x^k) = \nabla f(x^k)$. Updates (4) and (5) are equivalent as

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

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

$$\nabla_{\mathbf{S}_k} f(x^k) + \nabla^2_{\mathbf{S}_k} f(x^k) h^* + \frac{L_{\text{alg}}}{2}\|h^*\|_{x^k, \mathbf{S}_k} \nabla^2_{\mathbf{S}_k} f(x^k) h^* = 0, \tag{34}$$

which after rearranging is

$$h^* = -\left(1 + \frac{L_{\text{alg}}}{2}\|h^*\|_{x^k, \mathbf{S}_k}\right)^{-1} \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x^k), \tag{35}$$

thus solution of cubical regularization in local norms (6) has form of Newton method with stepsize $\alpha_k = \left(1 + \frac{L_{\text{alg}}}{2}\|h^*\|_{x^k, \mathbf{S}_k}\right)^{-1}$. We are left to show that this $\alpha_k$ is equivalent to (7).

Substitute $h^*$ from (35) to (34) and $\alpha_k = \left(1 + \frac{L_{\text{alg}}}{2}\|h^*\|_{x^k,\mathbf{S}_k}\right)^{-1}$ and then use $\nabla^2 f(x^k)[\nabla^2 f(x^k)]^\dagger \nabla f(x^k) = \nabla f(x^k)$, to get

$$\begin{aligned} 0 &= \nabla_{\mathbf{S}_k} f(x^k) + \nabla^2_{\mathbf{S}_k} f(x^k)\left(-\alpha_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x^k)\right) \\ &\quad + \frac{L_{\text{alg}}}{2}\left(\alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k}\right) \nabla^2_{\mathbf{S}_k} f(x^k)\left(-\alpha_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \nabla_{\mathbf{S}_k} f(x^k)\right) \\ &= \left(1 - \alpha_k - \frac{L_{\text{alg}}}{2}\alpha_k^2 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k}\right) \nabla_{\mathbf{S}_k} f(x^k). \end{aligned}$$

To conclude the proof, observe that stepsize $\alpha_k$ from (7) is set to be the positive root of polynomial $1 - \alpha_k - \frac{L_{\text{alg}}}{2}\alpha_k^2 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} = 0$. Because $\alpha_k$ corresponds to $h^*$ such that $\nabla_h T_{\mathbf{S}_k}\left(x^k, h\right)\big|_{h^*} = 0$, vector $h^*$ is the minimizer of the regularized model $T_{\mathbf{S}_k}\left(x^k, h\right)$ in (3). On the other hand, the equation (35) shows that $h^*$ has the form of Newton method with the stepsize (4).

This concludes the equivalence of (3), (4), and (5). □

### E.2 Proof of Lemma 1

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

□

### E.3 Proof of Lemma 2

*Proof.* We follow proof of (Hanzely et al., 2020, Lemma 5.2). Using definitions and the cyclic property of the matrix trace,

$$\begin{aligned} \mathbb{E}\left[\tau(\mathbf{S})\right] &= \mathbb{E}\left[\text{Tr}\left(\mathbf{I}^{\tau(\mathbf{S})}\right)\right] = \mathbb{E}\left[\text{Tr}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger\right)\right] \\ &= \mathbb{E}\left[\text{Tr}\left(\mathbf{P}_x\right)\right] = \text{Tr}\left(\frac{\tau}{d}\mathbf{I}^d\right) = \tau. \end{aligned}$$

□

### E.4 Proof of Lemma 3

*Proof.* Equalities in (12) and (13) follows from directly from definitions of the norms $\|\cdot\|_x^2, \|\cdot\|_x^{*2}$, projection $\mathbf{P}_x$ and properties of pseudoinverse $\mathbf{M} = \mathbf{M}\mathbf{M}^\dagger\mathbf{M}$, $\mathbf{M}^\dagger = \mathbf{M}^\dagger\mathbf{M}\mathbf{M}^\dagger$ (for $\mathbf{M} = \mathbf{S}^\top \nabla^2 f(x)\mathbf{S}$) and deterministicity of $h, g, \nabla^2 f(x)$.

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{P}_x h\|_x^2\right] &= \mathbb{E}\left[h^\top \mathbf{P}_x^\top \nabla^2 f(x)\mathbf{P}_x h\right] \\ &= \mathbb{E}\left[h^\top \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x) h\right] \\ &= \mathbb{E}\left[h^\top \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x) h\right] \\ &= \mathbb{E}\left[h^\top \nabla^2 f(x)\mathbf{P}_x h\right] \\ &= h^\top \nabla^2 f(x)\mathbb{E}\left[\mathbf{P}_x\right] h \\ &\overset{As.1}{=} \frac{\tau}{d}\|h\|_x^2, \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{P}_x^\top g\right\|_x^{*2}\right] &= \mathbb{E}\left[g^\top \mathbf{P}_x [\nabla^2 f(x)]^\dagger \mathbf{P}_x^\top g\right] \\
&= \mathbb{E}\left[g^\top \mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x)[\nabla^2 f(x)]^\dagger \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top g\right] \\
&= \mathbb{E}\left[g^\top \mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top g\right] \\
&= \mathbb{E}\left[g^\top \mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top g\right] \\
&= \mathbb{E}\left[g^\top \mathbf{S}\left(\mathbf{S}^\top \nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top \nabla^2 f(x)[\nabla^2 f(x)]^\dagger g\right] \qquad \text{if } g \in \text{Range}\left(\nabla^2 f(x)\right) \\
&= \mathbb{E}\left[g^\top \mathbf{P}_x [\nabla^2 f(x)]^\dagger g\right] \\
&= g^\top \mathbb{E}\left[\mathbf{P}_x\right] [\nabla^2 f(x)]^\dagger g \\
&\overset{As.1}{=} \frac{\tau}{d}\|g\|_x^{*2}.
\end{aligned}
$$

□

### E.5 Proof of Lemma 7

*Proof.* We have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{P}_x\right] &= \left[\nabla^2 f(x)\right]^{-1/2} \mathbb{E}_{\mathbf{M}\sim\tilde{\mathcal{D}}}\left[\mathbf{M}^\top \left(\mathbf{M}^\top \mathbf{M}\right)^\dagger \mathbf{M}\right] \left[\nabla^2 f(x)\right]^{1/2} \\
&= \left[\nabla^2 f(x)\right]^{-1/2} \frac{\tau}{d}\mathbf{I} \left[\nabla^2 f(x)\right]^{1/2} = \frac{\tau}{d}\mathbf{I}.
\end{aligned}
$$

□

### E.6 Proof of Lemma 4

*Proof.* For $h^k = x^{k+1} - x^k$, we can follow proof of (Hanzely et al., 2022, Lemma 10),

$$
\begin{aligned}
f(x^k) - f(x^{k+1}) &\overset{(16)}{\geq} -\left\langle \nabla_{\mathbf{S}} f(x^k), h^k \right\rangle - \frac{1}{2}\left\|h^k\right\|_{x^k,\mathbf{S}_k}^2 - \frac{L_{\text{alg}}}{6}\|h\|_{x^k,\mathbf{S}_k}^3 \\
&\overset{(8)}{=} \alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} - \frac{1}{2}\alpha_k^2 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} \\
&\qquad -\frac{L_{\text{alg}}}{6}\alpha_k^3 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*3} \\
&= \left(1 - \frac{1}{2}\alpha_k - \frac{L_{\text{alg}}}{6}\alpha_k^2 \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*}\right) \alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} \\
&\geq \frac{1}{2}\alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} \\
&\geq \frac{1}{2\max\left\{\sqrt{L_{\text{alg}}\|\nabla_{\mathbf{S}_k} f(x^k)\|_{x^k,\mathbf{S}_k}^{*}}, 2\right\}} \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2}.
\end{aligned}
$$

□

### E.7 Proof of Lemma 6

*Proof.* We bound norm of $\nabla_{\mathbf{S}} f(x^{k+1})$ using basic norm manipulation and triangle inequality as

$$\begin{aligned}
\left\|\nabla_{\mathbf{S}_k} f(x^{k+1})\right\|^*_{x^k,\mathbf{S}_k} &= \left\|\nabla_{\mathbf{S}_k} f(x^{k+1}) - \nabla^2_{\mathbf{S}_k} f(x^k)(x^{k+1}-x^k) - \alpha_k \nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&= \left\|\nabla_{\mathbf{S}_k} f(x^{k+1}) - \nabla_{\mathbf{S}_k} f(x^k) - \nabla^2_{\mathbf{S}_k} f(x^k)(x^{k+1}-x^k) + \right. \\
&\qquad \left. + (1-\alpha_k)\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&\le \left\|\nabla_{\mathbf{S}_k} f(x^{k+1}) - \nabla_{\mathbf{S}_k} f(x^k) - \nabla^2_{\mathbf{S}_k} f(x^k)(x^{k+1}-x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&\qquad + (1-\alpha_k)\left\|\nabla f(x^k)\right\|^*_{x^k,\mathbf{S}_k}.
\end{aligned}$$

Using $L_{\text{semi}}$–semi-strong self-concordance, we can continue

$$\begin{aligned}
\cdots &\le \left\|\nabla_{\mathbf{S}_k} f(x^{k+1}) - \nabla_{\mathbf{S}_k} f(x^k) - \nabla^2_{\mathbf{S}_k} f(x^k)(x^{k+1}-x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&\qquad + (1-\alpha_k)\left\|\nabla_{\mathbf{S}} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&\le \frac{L_{\text{semi}}}{2}\left\|x^{k+1}-x^k\right\|^2_{x^k,\mathbf{S}_k} + (1-\alpha_k)\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&= \frac{L_{\text{semi}}\alpha_k^2}{2}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*2}_{x^k,\mathbf{S}_k} + (1-\alpha_k)\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&= \left(\frac{L_{\text{alg}}\alpha_k^2}{2}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}} - \alpha_k + 1\right)\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \\
&\overset{(7)}{=} L_{\text{alg}}\alpha_k^2\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*2}_{x^k,\mathbf{S}_k}.
\end{aligned}$$

The last equality holds because of the choice of $\alpha_k$. □

### E.8 Technical lemmas

**Lemma 8** (Arithmetic mean – Geometric mean inequality)**.** *For $c \ge 0$ we have*

$$1 + c = \frac{1 + (1+2c)}{2} \overset{AG}{\ge} \sqrt{1+2c}. \tag{36}$$

**Lemma 9** (Jensen for square root)**.** *Function $f(x) = \sqrt{x}$ is concave, hence for $c \ge 0$ we have*

$$\frac{1}{\sqrt{2}}(\sqrt{c}+1) \le \sqrt{c+1} \qquad \le \sqrt{c}+1. \tag{37}$$

### E.9 Proof of Lemma 5

*Proof.* Denote

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

so that

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

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

$$f(x^{k+1}) \le \min_{h'\in\mathbb{R}^{\tau(\mathbf{S})}} T_{\mathbf{S}}\left(x^k, h'\right) = \min_{h''\in\mathbb{R}^d} \Omega_{\mathbf{S}}(x^k, h''),$$

and

$$\begin{aligned}
&\mathbb{E}\left[f(x^{k+1})\right] \\
&\quad\leq \quad \mathbb{E}\left[\Omega_{\mathbf{S}}(x^k,h)\right] \\
&\quad= \quad f(x^k)+\frac{\tau}{d}\left\langle\nabla f(x^k),h\right\rangle+\frac{1}{2}\mathbb{E}\left[\|\mathbf{P}_{x^k}h\|_{x^k}^2\right]+\mathbb{E}\left[\frac{L_{\text{alg}}}{6}\|\mathbf{P}_{x^k}h\|_{x^k}^3\right] \\
&\quad\overset{(12)}{\leq} \quad f(x^k)+\frac{\tau}{d}\left\langle\nabla f(x^k),h\right\rangle+\frac{\tau}{2d}\|h\|_{x^k}^2+\frac{L_{\text{alg}}}{6}\frac{\tau}{d}\|h\|_{x^k}^3 \\
&\quad\overset{(16)}{\leq} \quad f(x^k)+\frac{\tau}{d}\left(f(y)-f(x^k)+\frac{L_{\text{semi}}}{6}\left\|y-x^k\right\|_{x^k}^3\right)+\frac{L_{\text{alg}}}{6}\frac{\tau}{d}\|h\|_{x^k}^3.
\end{aligned}$$

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

### E.10 Proof of Theorem 2

*Proof.* Denote

$$A_0 \quad\overset{\text{def}}{=}\quad \frac{4}{3}\left(\frac{d}{\tau}\right)^3, \tag{38}$$

$$A_k \quad\overset{\text{def}}{=}\quad A_0+\sum_{t=1}^{k}t^2=A_0-1+\frac{k(k+1)(2k+1)}{6}\geq A_0+\frac{k^3}{3}, \tag{39}$$

$$\text{...consequently} \qquad \sum_{t=1}^{k}\frac{t^6}{A_t^2}\leq 9k, \tag{40}$$

$$\eta_t \quad\overset{\text{def}}{=}\quad \frac{d}{\tau}\frac{(t+1)^2}{A_{t+1}} \qquad \text{implying } 1-\frac{\tau}{d}\eta_t=\frac{A_t}{A_{t+1}}. \tag{41}$$

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

$$\eta_{t-1}\leq\frac{d}{\tau}\frac{t^2}{A_0+\frac{t^3}{3}}\leq\frac{d}{\tau}\sup_{t\in\mathbb{N}}\frac{t^2}{A_0+\frac{t^3}{3}}\leq\frac{d}{\tau}\sup_{\zeta>0}\frac{\zeta^2}{A_0+\frac{\zeta^3}{3}}=1 \tag{42}$$

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 $f$,

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1}|x^t)\right] \quad\leq\quad & \left(1-\frac{\tau}{d}\right)f(x^t)+\frac{\tau}{d}f^*\eta_t+\frac{\tau}{d}f(x^t)(1-\eta_t) \\
&+\frac{\tau}{d}\left(\frac{\max L_{\mathbf{S}}+L_{\text{semi}}}{6}\left\|x^t-x^*\right\|_{x^t}^3\eta_t^3\right).
\end{aligned}$$

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

$$\delta_{t+1}\leq\frac{A_t}{A_{t+1}}\delta_t+\frac{\max L_{\mathbf{S}}+L_{\text{semi}}}{6}\left\|x^t-x^*\right\|_{x^t}^3\left(\frac{d}{\tau}\right)^2\left(\frac{(t+1)^2}{A_{t+1}}\right)^3. \tag{43}$$

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

$$A_k\delta_k\leq A_0\delta_0+\frac{\max L_{\mathbf{S}}+L_{\text{semi}}}{6}\frac{d^2}{\tau^2}\sum_{t=0}^{k-1}\left\|x^t-x^*\right\|_{x^t}^3\frac{(t+1)^6}{A_{t+1}^2}. \tag{44}$$

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

$$\begin{aligned}
A_k\delta_k \quad&\leq\quad A_0\delta_0+\frac{\max L_{\mathbf{S}}+L_{\text{semi}}}{6}\frac{d^2}{\tau^2}D^3\sum_{t=1}^{k}\frac{t^6}{A_t^2} &(45)\\
&\leq\quad A_0\delta_0+\frac{\max L_{\mathbf{S}}+L_{\text{semi}}}{6}\frac{d^2}{\tau^2}D^3 9k, &(46)
\end{aligned}$$

and

$$\delta_k \leq \frac{A_0\delta_0}{A_k} + \frac{3(\max L_{\mathbf{S}} + L_{\text{semi}})d^2D^3k}{2\tau^2 A_k} \tag{47}$$

$$\leq \frac{3A_0\delta_0}{k^3} + \frac{9(\max L_{\mathbf{S}} + L_{\text{semi}})d^2D^3}{2\tau^2k^2}, \tag{48}$$

which concludes the proof. □

### E.11 Proof of Theorem 3

Before we start proof, we first state that for self-concordant functions (Definition 4) we can bound function value suboptimality by the norm of the gradient in the neighborhood of the solution.

**Proposition 3.** *(Hanzely et al., 2020, Lemma D.3) For any $\gamma > 0$ 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 $f : \mathbb{R}^d \to \mathbb{R}$, we can bound*

$$f(x^k) - f^* \leq \frac{1}{2}(1+\gamma)\left\|\nabla f(x^k)\right\|_{x^k}^{*2}. \tag{49}$$

*Proof of Theorem 3.* Note

$$\begin{aligned}
\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} &= \nabla f(x^k)^\top \mathbf{S}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top\nabla f(x^k) \\
&= \nabla f(x^k)^\top \mathbf{S}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top\nabla f(x^k) \\
&= \nabla f(x^k)^\top \mathbf{S}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top\nabla^2 f(x)[\nabla^2 f(x)]^\dagger\nabla^2 f(x)\cdot \\
&\qquad \cdot\mathbf{S}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top\nabla f(x^k) \\
&= \left\|\mathbf{P}_{x^k}^\top\nabla f(x^k)\right\|_{x^k}^{*2} \qquad (50) \\
&\overset{(11)}{\leq} \left\|\nabla f(x^k)\right\|_{x^k}^{*2},
\end{aligned}$$

and for $L_{\text{sc}}$–self-concordant function $f$ and $\gamma > 0$ in the neighborhood $x^k \in \left\{x : \|\nabla f(x)\|_x^* < \frac{2}{(1+\gamma^{-1})L_{\text{sc}}}\right\}$, we have

$$\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^* \leq \left\|\nabla f(x^k)\right\|_{x^k}^* < \frac{2}{(1+\gamma^{-1})L_{\text{sc}}}. \tag{51}$$

From Equation (15) we have

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{a_k}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} > \frac{1}{2b}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}_k}^{*2} \tag{52}$$

where

$$a_k \overset{\text{def}}{=} 2\max\left\{\sqrt{L_{\text{alg}}\|\nabla_{\mathbf{S}_k} f(x^k)\|_{x^k,\mathbf{S}_k}^*}, 2\right\} < 4\max\left\{\sqrt{\frac{L_{\text{alg}}}{2(1+\gamma^{-1})L_{\text{sc}}}}, 1\right\} \overset{\text{def}}{=} 2b$$

We can take the expectation and continue

$$\begin{aligned}
\mathbb{E}\left[f(x^k) - f(x^{k+1})\right] &\overset{(15)}{\geq} \mathbb{E}\left[\frac{1}{2b}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|_{x^k,\mathbf{S}}^{*2}\right] \\
&\overset{(50)}{=} \mathbb{E}\left[\frac{1}{2b}\left\|\mathbf{P}_{x^k}^\top\nabla f(x^k)\right\|_{x^k}^{*2}\right] \\
&\overset{(13)}{=} \frac{\tau}{2bd}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} \\
&\overset{(49)}{\geq} \frac{\tau}{bd(1+\gamma)}(f(x^k) - f^*).
\end{aligned}$$

Hence

$$\mathbb{E}\left[f(x^{k+1})-f^*)\right] \leq \left(1-\frac{\tau}{bd(1+\gamma)}\right)(f(x^k)-f^*),$$

and to finish the proof, we choose $\gamma=1$ and use tower property across iterates $x^0, x^1, \ldots, x^k$.

As semi-strong self-concordance is stronger than standard self-concordance (see Appendix D) and $L_{\text{semi}} \geq L_{\text{sc}}$, for simplicity of presentation we replace $L_{\text{semi}}$ by $L_{\text{sc}}$. □

### E.12 Towards proof of Theorem 4

**Proposition 4.** *(Gower et al., 2019, Equation (47)) Relative convexity* (19) *implies bound*

$$f^* \leq f(x^k) - \frac{1}{2\hat{\mu}}\left\|\nabla f(x^k)\right\|^{*2}_{x^k}. \tag{53}$$

**Proposition 5.** *Analogy to (Gower et al., 2019, Lemma 7) For $\mathbf{S}\sim\mathcal{D}$ satisfying conditions*

$$\text{Null}\left(\mathbf{S}^\top\nabla^2 f(x)\mathbf{S}\right) = \text{Null}\left(\mathbf{S}\right) \quad \textit{and} \quad \text{Range}\left(\nabla^2 f(x)\right) \subset \text{Range}\left(\mathbb{E}\left[\mathbf{S}_k\mathbf{S}_k^\top\right]\right), \tag{54}$$

*also, the exactness condition holds*

$$\text{Range}\left(\nabla^2 f(x)\right) = \text{Range}\left(\mathbb{E}\left[\hat{\mathbf{P}}_x\right]\right), \tag{55}$$

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

$$\rho(x) = \lambda^+_{min}\left(\mathbb{E}\left[\alpha_{x,\mathbf{S}}\mathbf{P}_x\right]\right) > 0 \tag{56}$$

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

**Lemma 10** (Stepsize bound)**.** *Stepsize $\alpha_k$ can be bounded as*

$$\alpha_k \leq \frac{\sqrt{2}}{\sqrt{L_{alg}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}}, \tag{57}$$

*and for $x^k$ far from solution, $\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \geq \frac{1}{L_{\mathbf{S}_k}}$ and $L_{alg} \geq \frac{9}{2}\sup_{\mathbf{S}} L_{\mathbf{S}}\hat{L}^2_{\mathbf{S}}$ holds $\alpha_k\hat{L}_{\mathbf{S}_k} \leq \frac{2}{3}$.*

*Proof of Lemma 10.* Denote $G_k \overset{\text{def}}{=} L_{\text{alg}}\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k}$. Using (37) with $c \leftarrow 2G > 0$ and

$$\alpha_k = \frac{-1+\sqrt{1+2G}}{G} \leq \frac{\sqrt{2G}}{G} = \frac{\sqrt{2}}{\sqrt{G}} = \frac{\sqrt{2}}{\sqrt{L_{\text{alg}}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}} \tag{58}$$

and

$$\begin{aligned}
\alpha_k\hat{L}_{\mathbf{S}_k} &\leq \frac{\sqrt{2}\hat{L}_{\mathbf{S}_k}}{\sqrt{L_{\text{alg}}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}} \\
&\leq \frac{\sqrt{2}\hat{L}_{\mathbf{S}_k}}{\sqrt{\frac{9}{2}L_{\mathbf{S}_k}\hat{L}^2_{\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}} \\
&\leq \frac{2}{3}\frac{1}{\sqrt{L_{\mathbf{S}_k}\|\nabla_{\mathbf{S}_k} f(x^k)\|^*_{x^k,\mathbf{S}_k}}} \leq \frac{2}{3}, \qquad \text{for } \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^*_{x^k,\mathbf{S}_k} \geq \frac{1}{L_{\mathbf{S}_k}}.
\end{aligned}$$

□

### E.12.1 Proof of Theorem 4

*Proof.* Replacing $x \leftarrow x^k$ and $h \leftarrow \alpha_k \mathbf{P}_{x^k}[\nabla^2 f(x^k)]^\dagger \nabla f(x^k)$ so that $x^{k+1} = x^k + \mathbf{S}h$ in (20) yields

$$f(x^{k+1}) \leq f(x^k) - \alpha_k \left(1 - \frac{1}{2}\hat{L}_{\mathbf{S}_k}\alpha_k\right) \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*2}_{x^k,\mathbf{S}_k} \tag{59}$$

$$\leq f(x^k) - \frac{2}{3}\alpha_k \left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*2}_{x^k,\mathbf{S}_k}. \tag{60}$$

In last step, we used that $\hat{L}_{\mathbf{S}_k}\alpha_k \leq \frac{2}{3}$ holds for $\left\|\nabla_{\mathbf{S}_k} f(x^k)\right\|^{*}_{x^k,\mathbf{S}_k} \geq \frac{1}{\hat{L}_{\mathbf{S}_k}}$ (Lemma 10). Next, we take an expectation over $x^k$ and use the definition of $\rho(x^k)$.

$$\mathbb{E}\left[f(x^{k+1})\right] \leq f(x^k) - \frac{2}{3}\left\|\nabla f(x^k)\right\|^2_{\mathbb{E}\left[\alpha_k \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \mathbf{S}_k^\top\right]} \tag{61}$$

$$= f(x^k) - \frac{2}{3}\nabla f(x^k)^\top \mathbb{E}\left[\alpha_k \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \mathbf{S}_k^\top\right] \nabla f(x^k) \tag{62}$$

$$= f(x^k) - \frac{2}{3}\nabla f(x^k)^\top \mathbb{E}\left[\alpha_k \mathbf{S}_k \left[\nabla^2_{\mathbf{S}_k} f(x^k)\right]^\dagger \mathbf{S}_k^\top\right] \nabla^2 f(x^k)[\nabla^2 f(x^k)]^\dagger \nabla f(x^k) \tag{63}$$

$$= f(x^k) - \frac{2}{3}\nabla f(x^k)^\top \mathbb{E}\left[\alpha_k \mathbf{P}_{x^k}\right] [\nabla^2 f(x^k)]^\dagger \nabla f(x^k) \tag{64}$$

$$\leq f(x^k) - \frac{2}{3}\rho(x^k)\nabla f(x^k)^\top [\nabla^2 f(x^k)]^\dagger \nabla f(x^k) \tag{65}$$

$$= f(x^k) - \frac{2}{3}\rho(x^k)\left\|\nabla f(x^k)\right\|^{*2}_{x^k} \tag{66}$$

$$\overset{(53)}{\leq} f(x^k) - \frac{4}{3}\rho(x^k)\hat{\mu}\left(f(x^k) - f^*\right). \tag{67}$$

Now $\rho(x^k) \geq \rho$, and $\rho$ is bounded in Proposition 5. Rearranging and subtracting $f^*$ gives

$$\mathbb{E}\left[f(x^{k+1}) - f^*\right] \leq \left(1 - \frac{4}{3}\rho\hat{\mu}\right)(f(x^k) - f^*), \tag{68}$$

which after using tower property across all iterates yields the statement. □