

Fast Linear Convergence of Randomized BFGS

Dmitry Kovalev¹, Robert M. Gower², Peter Richtárik¹, and Alexander Rogozin^{1,3}

¹King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

²Télécom ParisTech, Paris, France

³Moscow Institute of Physics and Technology, Dolgoprudny, Russia

February 9, 2020
(Revised: February 26, 2020)

Abstract

Since the late 1950's when quasi-Newton methods first appeared, they have become one of the most widely used and efficient algorithmic paradigms for unconstrained optimization. Despite their immense practical success, there is little theory that shows why these methods are so efficient. We provide a semi-local rate of convergence for the randomized BFGS method which can be significantly better than that of gradient descent, finally giving theoretical evidence supporting the superior empirical performance of the method.

Contents

1	Introduction	2
1.1	Background	3
1.2	Notation and definitions	3
2	Convergence Results	4
2.1	Local linear convergence for self-concordant functions	4
2.2	Local linear convergence for smooth and strongly convex functions	5
2.3	Superlinear convergence	6
3	Examples and Applications	6
3.1	Invertible \mathbf{S}	6
3.2	One column \mathbf{S}	6
3.3	Generalized linear models	7
3.4	Linear programs with box constraints	8
4	Numerical Experiments	9
4.1	Synthetic quadratic problem	9
4.2	Classification problem on real data	10
5	Conclusion, Consequences and Future Work	14
A	Proof of Theorem 2.1	17
A.1	Properties of self-concordant functions	17
A.2	The distance of the iterates	18
A.3	The distance of the quasi-Newton matrix	19
A.4	Detailed proof of Theorem 2.1	21
B	Proof of Theorem 2.5	21
B.1	Getting ready for the proof	21
B.2	The Proof	23
C	Proof of Theorem 2.6	25

1 Introduction

In this paper we consider the optimization problem

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

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a twice continuously differentiable function. While the BFGS method (Broyden, 1967; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) is one of the most efficient and celebrated algorithms for solving (1), a clear theoretical justification for its success, or the success of *any* quasi-Newton method, has been elusive. We do know, however, that the quasi-Newton methods converge Q -superlinearly (Powell, 1971), and this is often pointed to as the justification for their success. Yet this superlinear convergence occurs only asymptotically, in an arbitrarily small ball around the solution, and at an unknown superlinear rate. In particular, this superlinear rate could be arbitrary close to linear.

In this work we provide *the first meaningful convergence rate of (a randomized variant of) BFGS*. By meaningful, we mean a rate that can be faster than the rate of gradient descent, thus giving much stronger support to the practical success of the BFGS method than what has been available so far. Furthermore, for our results to hold, we only need to assume f to be self-concordant.

Let \mathbb{S}^d denote the set of $d \times d$ symmetric matrices, and let

$$\mathbf{H}_x := \nabla^2 f(x)$$

be the Hessian matrix of f evaluated at x . For $x = x_k$ we will further abbreviate $\mathbf{H}_k := \mathbf{H}_{x_k}$. In this paper we consider the randomized BFGS update introduced by Gower & Richtárik (2017); Gower et al. (2016), given by the formula

$$\text{BFGS}(\mathbf{B}, \mathbf{H}, \mathbf{S}) := \mathbf{G} + (\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{B}(\mathbf{I} - \mathbf{H}\mathbf{G}), \quad (2)$$

where

$$\mathbf{G} = \mathbf{G}(\mathbf{H}, \mathbf{S}) := \mathbf{S}(\mathbf{S}^\top \mathbf{H} \mathbf{S})^{-1} \mathbf{S}^\top,$$

$\mathbf{S} \in \mathbb{R}^{d \times \tau}$ is a (typically thin) random matrix with $\tau \ll d$ columns, $\mathbf{H} \in \mathbb{S}^d$ is a target Hessian matrix and $\mathbf{B} \in \mathbb{S}^d$ is the current estimate of the inverse Hessian matrix. We use this BFGS update in the randomized BFGS method stated as Algorithm 1.

Algorithm 1 RBFGS

- 1: **Parameters:** $x_0 \in \mathbb{R}^d$, $\mathbf{B}_0 \in \mathbb{S}^d$, distribution \mathcal{D} over vectors from \mathbb{R}^d
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $\mathbf{H}_k = \nabla^2 f(x_k)$
 - 4: Sample random $\mathbf{S}_k \sim \mathcal{D}$
 - 5: $\mathbf{B}_{k+1} = \text{BFGS}(\mathbf{B}_k, \mathbf{H}_k, \mathbf{S}_k)$
 - 6: $x_+ = x_k - \mathbf{B}_k \nabla f(x_k)$
 - 7: **Monotonic option:**
 - 8: $x_{k+1} = \arg \min \{f(x_+), f(x_k)\}$
 - 9: **end for**
-

In order to update our estimate of the inverse Hessian \mathbf{H}_k^{-1} , we use a *random linear* measurement $\mathbf{H}_k \mathbf{S}_k$ of the true Hessian \mathbf{H}_k . Random linear transformations are often alternatively called *sketches*, and we adopt this terminology. We thus refer to the random matrix $\mathbf{S}_k \in \mathbb{R}^{d \times \tau}$ by the name *sketching matrix*, and to the product $\mathbf{H}_k \mathbf{S}_k$ by the name *Hessian sketch*. Note that the sketch can be computed by doing τ directional derivatives of the gradient since

$$\left. \frac{d}{dt} \nabla f(x_k + \mathbf{S}_k t) \right|_{t=0} = \nabla^2 f(x_k) \mathbf{S}_k = \mathbf{H}_k \mathbf{S}_k. \quad (3)$$

Our forthcoming theoretical results hold virtually any distribution \mathcal{D} of sketching matrices.

Based on the intuition our general theory provides, in the case of generalized linear models (see Section 3.3) we also develop a particular distribution \mathcal{D} based on the singular value decomposition for. In Section 4 we show that our new sketching matrix is well suited to highly ill-conditioned problems. We also compare our new sketch against other standard sketches and the original BFGS method in a series of numerical experiments.

1.1 Background

We now recall the background on the convergence of BFGS, some modern randomized BFGS variants and their convergence results, and also techniques surrounding self-concordance.

Classic convergence results. The BFGS method was first shown to converge locally and superlinearly over 50 years ago (Powell, 1971). This proof was later extended to include a larger family of quasi-Newton methods by Broyden et al. (1973). Much more recently, Gao & Goldfarb (2016) showed that the block BFGS method (Gower et al., 2016) also converges superlinearly (and without the need for line search).

Modern randomized variants of BFGS. Byrd et al. (2016) proposed the SQN method which uses a single Hessian-vector product in the BFGS update, as opposed to using differences of stochastic gradients. Subsequently, Moritz et al. (2015) propose combining SQN with the variance reduced gradient method SVRG (Johnson & Zhang, 2013), and provide a global linear convergence rate for the resulting method. Gower et al. (2016) extend the SQN method to allow for sketching matrices (3), and also provide an improved linear convergence rate. However, their rate is still orders of magnitude worse than the rate of convergence of gradient descent. Very recently, Meng et al. (2019) give a global convergence rate for a stochastic variant of the BFGS method combined with stochastic gradient descent. The reason that the rates of convergence in all of these previous works are significantly inferior to that of gradient descent, is that they do not factor in the contribution of the quasi-Newton matrix towards the convergence. Instead, the previous analysis focus on obtaining bounds on the estimated inverse Hessian

$$c\mathbf{I} \preceq \mathbf{B}_k \preceq C\mathbf{I}, \quad \text{where } 0 < c < C, \quad (4)$$

and then use this to “bound away” the contribution of the quasi-Newton method, for which a cost is paid. Following this step, the analysis follows verbatim the standard analysis of gradient-based methods, albeit with the added burden that the constants c and C (4) bring.

Self-concordance. It was a revolution¹ in the optimization community when, in the late 70’s and early 80’s, it was shown that all convex optimization problems could be solved in polynomial time by interior point methods (Nesterov & Nemirovskii, 1987). A key concept that facilitated this revolution was that of self-concordance, which describes a large class of convex functions whose second order derivative could be naturally “controlled” by the third derivative. We rely on self-concordant functions in this paper in Theorem 2.1, and in the proof, show how the convergence of the BFGS method can leverage self-concordance.

1.2 Notation and definitions

Both quasi-Newton methods and self-concordance are defined using weighted norms. For every $\mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{H} \in \mathbb{S}^d$ we write

$$\|\mathbf{W}\|_{F(\mathbf{H})}^2 := \text{Trace}(\mathbf{H}\mathbf{W}\mathbf{H}\mathbf{W}^\top) = \left\| \mathbf{H}^{1/2}\mathbf{W}\mathbf{H}^{1/2} \right\|_F^2$$

to denote a weighted Frobenius norm. Let $\|v\|_x^2 := \langle \mathbf{H}_x v, v \rangle$ be the *local inner product* (Renegar, 2001). To further abbreviate our notation, we will use $\|v\|_* := \|v\|_{x_*}$. Let

$$\mathcal{B}_x^\delta := \{y \in \mathbb{R}^d : \|y - x\|_x < \delta\}$$

be the ball of size $\delta > 0$ around x under the local norm. Using the local inner product, we can now state the definition of the key concept of self-concordance.

Definition 1.1. A functional $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is self-concordant if for all $x \in \text{dom}_f$ its Hessian is positive definite, and for each $y \in \mathcal{B}_x^1$ we have

$$1 - \|y - x\|_x \leq \frac{\|v\|_x}{\|v\|_y} \leq \frac{1}{1 - \|y - x\|_x} \quad (5)$$

¹A precursor of this result made the front page of 1979 New York Times entitled “Soviet Discovery Rocks World of Mathematics”. It also made the headlines of the Guardian, with a humorously incorrect title “Soviet Answer to Travelling Salesman Problem” (Floudas & Pardalos, 2006).

for all $v \neq 0$.

2 Convergence Results

In this section we establish our main convergence results: local linear convergence for self-concordant functions, local linear convergence for smooth and strongly convex functions and superlinear convergence.

2.1 Local linear convergence for self-concordant functions

Under the assumption of self-concordance only, we now prove semi-local linear convergence of the randomized BFGS method in the following theorem.

Theorem 2.1. Let

$$\rho := \sup_{x \in \mathbb{R}^d} \lambda_{\min} \left(\mathbb{E} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right) \quad (6)$$

and consider the Lyapunov function

$$\Phi_\sigma^k := \sigma \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_*)}^2 + \|x_k - x_*\|_* , \quad (7)$$

where $\sigma = \frac{3}{7\rho}$. If f is self-concordant and

$$\|x_0 - x_*\|_* \leq \frac{1}{2} \min \left\{ \frac{1}{2}, 1 - \sqrt{\frac{1-\rho}{1-\frac{2\rho}{3}}}, \frac{4-2\rho}{9\rho d + 10} \right\} \quad (8)$$

then Algorithm 1 converges linearly according to

$$\mathbb{E} [\Phi_\sigma^{k+1}] = \left(1 - \frac{\rho}{2} \right) \mathbb{E} [\Phi_\sigma^k] . \quad (9)$$

Unrolling this recurrence, we get

$$\mathbb{E} [\|x_k - x_*\|_*] \leq \left(1 - \frac{\rho}{2} \right)^k \Phi_\sigma^0 . \quad (10)$$

We now provide a few key insights and a brief outline of the proof technique of Theorem 2.1; the complete proof is given in Section A in the Appendix.

Our starting point is the work of Gower & Richtárik (2017) who studied randomized algorithms for inverting a fixed invertible² matrix \mathbf{H} . Their methods can be interpreted as as a randomized, non-adaptive (and possibly block) variants of classical quasi-Newton matrix inversion formulas. In particular, they studied randomized BFGS updates for inverting $\mathbf{H} \succ 0$ and showed that the sequence of random matrices given by the randomized BFGS rule

$$\mathbf{B}_{k+1} = \text{BFGS}(\mathbf{B}_k, \mathbf{H}, \mathbf{S}_k)$$

converges linearly to \mathbf{H}^{-1} in mean square.

In this work, we face the additional challenge of a moving target \mathbf{H} . Indeed, in our setting, \mathbf{H} is not fixed, but is set to $\mathbf{H} = \mathbf{H}_k = \nabla^2 f(x_k)$ in iteration k , and hence changes from iteration to iteration. If the Hessian \mathbf{H}_k changes too fast, then there is no reason why the randomized BFGS update should be able to “catch up” with the change, and be any good at maintaining a good estimate of the inverse of \mathbf{H}_k throughout the iterations, let alone improve the estimate and converge to $(\nabla^2 f(x_*))^{-1}$ as $k \rightarrow \infty$.

This is the place where self-concordance comes to our aid. Indeed, self-concordance provides control over how fast the Hessian, and the inverse Hessian, changes as the point x_k changes; see Lemma A.1 in the Appendix. By carefully combining these intuitions, we are able to control how much \mathbf{B}_k deviates from the Hessian \mathbf{H}_k ; see Lemma A.5. Yet stringing this result into an induction argument is difficult due to the changing local norm. Fortunately, self-concordance allows us to “change the norm” to another point (5). Using this defining property of

²An extension to the computation of the Moore-Penrose pseudoinverse of a general rectangular matrix was developed in (Gower & Richtárik, 2016).

self-concordance, we state all of our results with respect to the local norm at the optimal point. This allows for a more straightforward induction argument and the resulting recurrence in (9).

2.2 Local linear convergence for smooth and strongly convex functions

Here we prove that BFGS converges with same rate of convergence as in Theorem 2.1, but under an alternative (though also common) set of assumptions. We also now use the Monotonic option on line 8 of Algorithm 1. This enforces that the function values $f(x_k)$ are monotonically decreasing. That is, let

$$\mathcal{Q} := \{x : f(x) \leq f(x_0)\}.$$

By (1) we have that $x_k \in \mathcal{Q}$ for all k .

Instead of self-concordance, we now rely on strong convexity, Lipschitz gradients and Hessians.

Assumption 2.2. Function f is μ -strongly convex. That is,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \quad (11)$$

for all $x, y \in \mathbb{R}^d$.

Assumption 2.3. Function f has L_1 -Lipschitz gradient. That is,

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L_1 \|y - x\|_2 \quad (12)$$

for all $x, y \in \mathbb{R}^d$.

Assumption 2.4. Function f has L_2 -Lipschitz Hessian. That is,

$$\|\mathbf{H}_y - \mathbf{H}_x\|_2 \leq L_2 \|y - x\|_2 \quad (13)$$

for all $x, y \in \mathbb{R}^d$.

Under these assumptions we have the following theorem.

Theorem 2.5. Let $f(x_0) - f(x_*) \leq \mathcal{F}$, where \mathcal{F} is defined as follows:

$$\mathcal{F} := \frac{1}{4} \left[\frac{\sqrt{2}L_1L_2}{\mu^2} + \frac{32\sqrt{2}dL_1^{5/2}L_2}{\rho\mu^4} \right]^{-2}. \quad (14)$$

If Assumptions 2.2, 2.3 and 2.4 hold, then

$$\mathbb{E} [\Psi_{k+1}] \leq \left(1 - \frac{\rho}{2}\right) \mathbb{E} [\Psi_k], \quad (15)$$

where

$$\rho := \sup_{x \in \mathcal{Q}} \lambda_{\min} \left(\mathbb{E} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right). \quad (16)$$

A disadvantage of using Assumptions 2.2, 2.3, and 2.4, as we do in Theorem 2.5, as compared to self-concordance, as we do in Theorem 2.1, is that now the region of local convergence (14) depends on constants which can be hard, or almost numerically impossible (such as μ), to compute. As such, it would be very difficult to use Theorem 2.5 together with globalization strategies to develop a global linear convergence. In contrast, we only need to have a bound on ρ to compute the region of convergence in Theorem 2.5 (see (8)). We give some insights into bounding ρ in Section 3.

2.3 Superlinear convergence

For completion, we also prove the superlinear convergence of Algorithm 1 with high probability.

Theorem 2.6. $\sqrt{f(x_k) - f(x_*)}$ converges to 0 at superlinear rate with probability 1.

Proof. The proof follows by combining Lemma C.1 and Lemma C.2. \square

3 Examples and Applications

To gain some insight into the consequences of Theorems 2.5 and 2.1, we first examine two extreme cases for the choice of the sketching matrix: i) \mathbf{S} is invertible (full curvature), ii) \mathbf{S} has only one column. We then develop a more involved SVD sketch, and apply it to generalized linear models and linear programming.

3.1 Invertible \mathbf{S}

When the sketching matrix \mathbf{S} is invertible with probability one, then $\rho = 1$ and consequently by (8) we have that if

$$\|x_0 - x_*\|_* \leq \min \left\{ \frac{1}{4}, \frac{1}{9d+10} \right\} = \frac{1}{9d+10},$$

then

$$\|x_k - x_*\|_* \leq \left(\frac{1}{2} \right)^k \Phi_\sigma^0.$$

This is not surprising. Indeed, given that when \mathbf{S} is invertible, it is not hard to show that $\mathbf{B}_k = \mathbf{H}_k^{-1}$; that is, the quasi-Newton matrix equals the inverse Hessian. Consequently, Algorithm 1 is equivalent to Newton's method.

3.2 One column \mathbf{S}

On the other extreme, if we choose the sketch matrix to have a single column only, $\mathbf{S} = s \in \mathbb{R}^d$, then ρ can be much smaller than 1 and will also depend on the spectrum of the Hessian.

Corollary 3.1 (Single column sketches). Let $0 \prec \mathbf{U} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix such that $\mathbf{H}_x \preceq \mathbf{U}$, $\forall x \in \mathbb{R}^d$. Let $\mathbf{D} = [d_1, \dots, d_n] \in \mathbb{R}^{n \times n}$ be a given invertible matrix such that $d_i^\top \mathbf{H}_x d_i \neq 0$ for all $x \in \mathbb{R}^d$ and $i = 1, \dots, n$. Furthermore, suppose that f is convex and the Hessian matrix is Lipschitz. If we sample according to

$$\mathbb{P}(\mathbf{S}_k = d_i) = p_i := \frac{d_i^\top \mathbf{U} d_i}{\text{Trace}(\mathbf{D}^\top \mathbf{U} \mathbf{D})},$$

then, under the assumptions of Theorem 2.5, Algorithm 1 converges at a rate of at least

$$\rho \geq \min_{x \in \mathcal{Q}} \frac{\lambda_{\min}^+(\mathbf{H}_x^{1/2} \mathbf{D} \mathbf{D}^\top \mathbf{H}_x^{1/2})}{\text{Trace}(\mathbf{D}^\top \mathbf{U} \mathbf{D})}. \quad (17)$$

Proof. This quantity ρ also appears as the convergence rate of the randomized subspace Newton method in (Gower et al., 2019). In particular, it was shown in Corollary 1 in (Gower et al., 2019) that under these assumptions, (17) holds. \square

Thus, by carefully choosing the sketching matrix \mathbf{S} , we can see that the rate of convergence ρ is bounded below as in (17). Since in this example we have not assumed the spectrum of the Hessian to be bounded from below, ρ still depends on \mathbf{H}_x in (17).

3.3 Generalized linear models

We now show that in situations where the Hessian matrix has a bounded spectrum, we can choose a sketching matrix so that ρ essentially does *not depend* on the spectra or ill-conditioning of the Hessian matrix. This leads to a more precise bound on the rate of convergence ρ that highlights how it can be arbitrarily bigger (leading to faster rate) than the rate of convergence of gradient descent. This also extends to accelerated gradient descent.

Definition 3.2. Let $0 \leq \ell \leq u$. Let $\phi_i : \mathbb{R} \mapsto \mathbb{R}_+$ be a twice differentiable function such that

$$\phi_i''(t) \in [\ell, u], \quad \text{and} \quad \phi_i'''(t) \leq U \quad \text{for } i = 1, \dots, n. \quad (18)$$

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

$$f(x) = \frac{1}{n} \sum_{i=1}^n \phi_i(\langle a_i, x \rangle). \quad (19)$$

Proposition 3.3. Let \mathbf{A} have full row rank. Let f be given by (19). Let $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the reduced singular value decomposition of \mathbf{A} with $\mathbf{U} \in \mathbb{R}^{d \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{n \times d}$. Consider a sketching matrix such that

$$\mathbb{P}[\mathbf{S} = \mathbf{U}\Sigma^{-1}e_i] = \frac{1}{d}, \quad \text{for } i = 1, \dots, n. \quad (20)$$

It follows that ρ given in (6) is bounded by

$$1 - \frac{\rho}{2} \leq 1 - \frac{1}{2} \frac{\ell}{u} \frac{1}{d}. \quad (21)$$

Proof. Since \mathbf{A} has full row rank we have that Σ is invertible, \mathbf{U} is orthonormal and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$. Consequently $\mathbf{A}^\dagger = \mathbf{V}\Sigma^{-1}\mathbf{U}^\top$. The Hessian of f is thus given by

$$\mathbf{H}_x = \frac{1}{n} \mathbf{A} \Phi''(\mathbf{A}^\top x) \mathbf{A}^\top, \quad (22)$$

where $\Phi''(\mathbf{A}^\top x) = \text{Diag}(\phi_1''(\langle a_1, x \rangle), \dots, \phi_n''(\langle a_n, x \rangle))$. It is now not hard to show that

$$\frac{\ell \sigma_{\min}(\mathbf{A})^2}{n} \mathbf{I} \preceq \mathbf{H}_x \preceq \frac{u \sigma_{\max}(\mathbf{A})^2}{n} \mathbf{I}, \quad (23)$$

and thus f is strongly convex with Lipschitz gradients. Furthermore, $\phi_i'''(t) \leq U$ guarantees that the Hessian is Lipschitz. Consequently, Theorem 2.5 holds. It now remains to bound ρ . With probability $\frac{1}{d}$, we have

$$\begin{aligned} \mathbf{S}^\top \mathbf{H}_x \mathbf{S} &= \frac{1}{n} e_i^\top \Sigma^{-1} \mathbf{U}^\top \mathbf{U} \Sigma \mathbf{V}^\top \Phi''(\mathbf{A}^\top x) \mathbf{V} \Sigma \mathbf{U}^\top \mathbf{U} \Sigma^{-1} e_i \\ &= \frac{1}{n} e_i^\top \mathbf{V}^\top \Phi''(\mathbf{A}^\top x) \mathbf{V} e_i. \end{aligned} \quad (24)$$

Now, since $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, we get

$$\ell \mathbf{I} \preceq \mathbf{V}^\top \Phi''(\mathbf{A}^\top x) \mathbf{V} \preceq u \mathbf{I}. \quad (25)$$

Consequently,

$$\begin{aligned} \rho = \lambda_{\min} \left(\mathbb{E} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right) &= \lambda_{\min} \left(\mathbb{E} \left[\frac{\mathbf{H}_x^{1/2} \mathbf{S} \mathbf{S}^\top \mathbf{H}_x^{1/2}}{\frac{1}{n} e_i^\top \mathbf{V}^\top \Phi''(\mathbf{A}^\top x) \mathbf{V} e_i} \right] \right) \\ &\geq \frac{n}{u} \lambda_{\min} \left(\mathbf{H}_x^{1/2} \mathbb{E} [\mathbf{S} \mathbf{S}^\top] \mathbf{H}_x^{1/2} \right) \\ &= \frac{n}{u} \lambda_{\min} \left(\mathbf{H}_x^{1/2} \mathbf{U} \Sigma^{-1} \mathbb{E} [e_i e_i^\top] \Sigma^{-1} \mathbf{U}^\top \mathbf{H}_x^{1/2} \right) \\ &= \frac{n}{ud} \lambda_{\min} (\Sigma^{-1} \mathbf{U}^\top \mathbf{H}_x \mathbf{U} \Sigma^{-1}) \\ &= \frac{1}{ud} \lambda_{\min} (\mathbf{V}^\top \Phi''(\mathbf{A}^\top x) \mathbf{V}) \geq \frac{\ell}{u} \frac{1}{d}. \end{aligned}$$

□

We refer to the sketch in (20) as the SVD-sketch. To contrast the bound on ρ given in (30), let us compare it to the rate of convergence of gradient descent. Gradient descent converges at a rate of

$$\rho_{GD} := 1 - \frac{\max_x \lambda_{\min}(\mathbf{H}_x)}{\min_x \lambda_{\max}(\mathbf{H}_x)} \leq 1 - \frac{\ell}{u} \frac{\sigma_{\min}(\mathbf{A})^2}{\sigma_{\max}(\mathbf{A})^2},$$

and is thus at the mercy of the condition number

$$\kappa := \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

of the matrix \mathbf{A} . To highlight this, consider the extreme case where \mathbf{A} is a square $d \times d$ Hilbert matrix (Hilbert, 1894). In this case

$$\kappa = O\left((1 + \sqrt{2})^{4d}\right),$$

and thus ρ_{GD} grows exponentially with d while ρ grows at most linearly due to (30). Even the rate of convergence of accelerated gradient descent (Nesterov, 1998) would grow exponentially with an exponent of $2d$ in this case.

Though in practice it is well known that the convergence of gradient descent is far more sensitive to ill conditioning as compared to the convergence of the BFGS method, to the best of our knowledge, this is the first clear theoretical justification to support this observation.

3.4 Linear programs with box constraints

The original problem that motivated the introduction of self-concordant functions was linear programming,

$$\begin{aligned} & \min_{x \in \mathbb{R}^d} \langle c, x \rangle \\ & \text{subject to } \mathbf{A}x \leq b, \end{aligned} \tag{26}$$

where $c \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ are the given data. Interior point methods and barrier methods for solving (26) are based on solving a sequence of problems such as

$$f(x) = \lambda \langle c, x \rangle - \sum_{i=1}^n \log(b_i - \langle a_i, x \rangle), \tag{27}$$

where $\lambda > 0$ is the barrier parameter, see (Renegar, 2001). The objective function in (27) is self-concordant. Furthermore, the logarithmic barrier enforces that the constraint $\mathbf{A}x < b$ hold.

The Hessian of (27) is given by

$$\nabla^2 f(x) = \mathbf{A}^\top \text{Diag} \left(\frac{1}{(b_i - \langle a_i, x \rangle)^2} \right) \mathbf{A}. \tag{28}$$

Let x^k be the iterates of Algorithm 1 applied to minimizing (27) with the monotonic option on line 8.

Proposition 3.4. Let \mathbf{A} have full row rank, f be given by (27), and $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the reduced singular value decomposition of \mathbf{A} with $\mathbf{U} \in \mathbb{R}^{d \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{n \times d}$. Consider the SVD-sketch given in (20). Let

$$\begin{aligned} u &= \frac{1}{\min_{i=1, \dots, n} \min_{x: \mathbf{A}x < b, x \in \mathcal{Q}} (b_i - \langle a_i, x \rangle)^2} \\ \ell &= \frac{1}{\max_{i=1, \dots, n} \max_{x: \mathbf{A}x < b, x \in \mathcal{Q}} (b_i - \langle a_i, x \rangle)^2}. \end{aligned} \tag{29}$$

It follows that ρ given in (6) is bounded by

$$1 - \frac{\rho}{2} \leq 1 - \frac{1}{2} \frac{\ell}{u} \frac{1}{d}. \quad (30)$$

Proof. The proof follows verbatim the proof of Proposition 3.3. \square

As the barrier parameter λ becomes larger, the minimizer of (27) will get closer to the boundary of the set $\mathbf{A}x \leq b$. Consequently, the Hessian (28) becomes increasing ill-conditioned, which is a well known issue with log-barrier methods. This in turn also implies that the constants u and ℓ in (29) will increase and decrease, respectively, as λ grows. Thus like all methods for solving the log-barrier problem (27), this ill-conditioning needs to be treated with care.

4 Numerical Experiments

We run numerical tests for the Randomized BFGS algorithm and compare three different distributions of \mathbf{S} :

1. **gauss**: All entries of \mathbf{S} are generated independently and have standard Gaussian distribution.
2. **svd**: SVD-sketch from Proposition 3.3, i.e., for singular value decomposition of $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ take

$$\mathbb{P}[\mathbf{S} = \mathbf{U}\Sigma^{-1}e_i] = \frac{1}{d} \quad \text{for } i = 1, \dots, n.$$

3. **coord**: Coordinate sketch, i.e., $\mathbb{P}[\mathbf{S} = e_i] = \frac{1}{d}$.

We also experiment with the sketch size, i.e. the factor τ for the sketching matrix \mathbf{S} . We use **distr** $_\tau$ in the legends of our plots to denote which distribution (**distr**) and which sketch size τ was used. For example, **gauss**₁₀ corresponds to using a Gaussian sketch with sketch size 10. Finally, we compare the performance of our method to the classical BFGS and accelerated Nesterov method both on synthetic and real data.

For a fair comparison across algorithms, we use the Wolfe line search Wolfe (1971) to determine the stepsizes for the Randomized BFGS, classical BFGS and Nesterov methods.

4.1 Synthetic quadratic problem

Consider a quadratic problem

$$\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\mathbf{A}x\|_2^2 = \frac{1}{2} x^\top (\mathbf{A}^\top \mathbf{A}) x \right\},$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a Hilbert matrix Hilbert (1894) defined as

$$[\mathbf{A}]_{ij} = \frac{1}{i + j - 1} \quad \text{for } i = 1, \dots, n.$$

This problem is interesting due to the large condition number of \mathbf{A} . For each distribution, we experiment with different sketch sizes and choose the one that gives the best results in terms of computational time. Another possibility is to compare the number of iterations. However, due to the fact that the complexity of each Algorithm 1 depends on τ , comparing the iteration number does not illustrate the algorithms performance well.

When using the **svd** sketch, we only include singular values above tolerance 10^{-8} . This is done to improve numerical stability. As a benchmark, we also compare to the classical BFGS method and accelerated Nesterov gradient descent, see Figure 1.

We observe in Figure 1 that for this particular quadratic problem small sketch sizes (10 or below) perform better. We note that **svd** distribution yielded the best results. This is to be expected, since according to Proposition 3.3, the rate of convergence when using the svd sketch is essentially invariant to ill-conditioning.

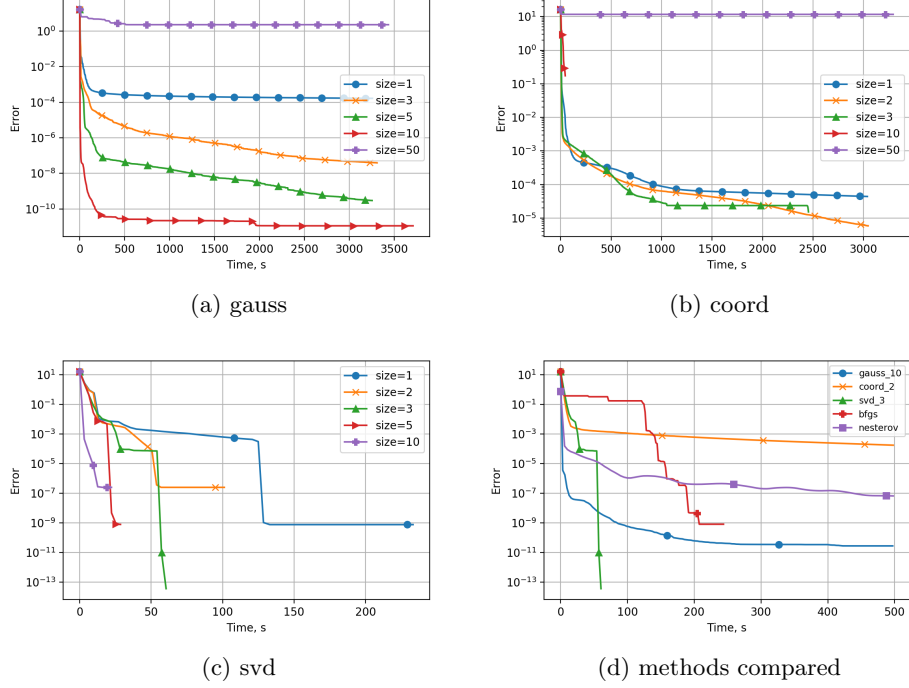


Figure 1: Hilbert matrix of dimension $10,000 \times 10,000$.

4.2 Classification problem on real data

Now consider the logistic regression problem with L2 regularizer given by

$$f(x) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i \langle a_i, x \rangle)) + \frac{\lambda}{2} \|x\|_2^2, \quad (31)$$

where $a_1, \dots, a_n \in \mathbb{R}^d$ are the data points, $b_1, \dots, b_n \in \{-1, 1\}$ are the class labels and $\lambda > 0$ is a regularization coefficient. In our experiments, we used LIBSVM datasets Chang & Lin (2011). We set the regularization parameter $\lambda \sim 10^{-3}L$, where L denotes the smoothness constant of logistic loss without regularizer. Similar to Section 4.1, we experiment with different sketch sizes and choose ones which perform better. The experiments show that a good general strategy of choosing sketch size is $\tau \sim \sqrt{d}$, see Figures 2, 3, 4, 7 and 8.

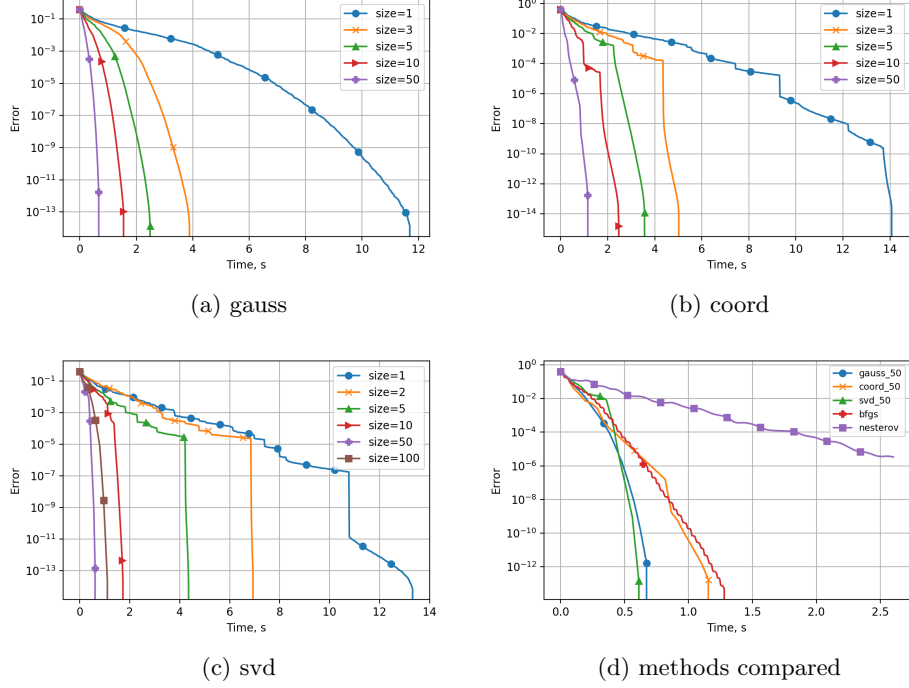


Figure 3: a9a, $\lambda = 10^{-3}$, $n = 29,159$, $d = 123$

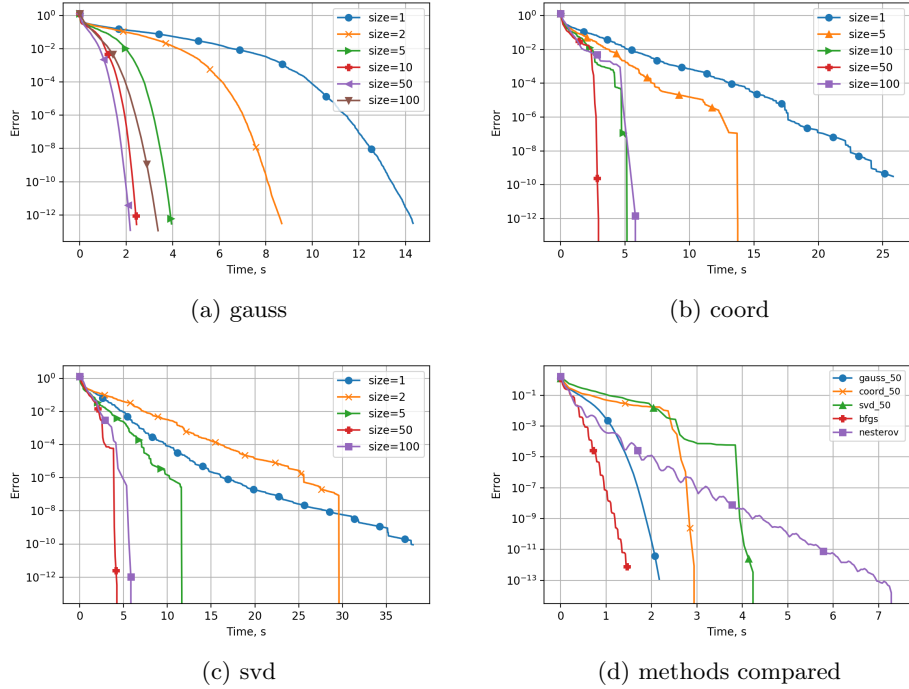
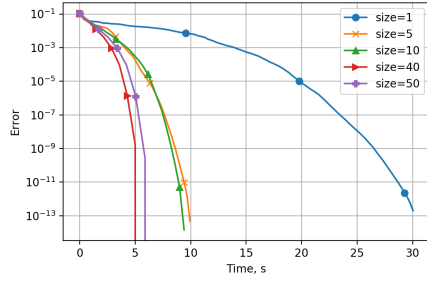


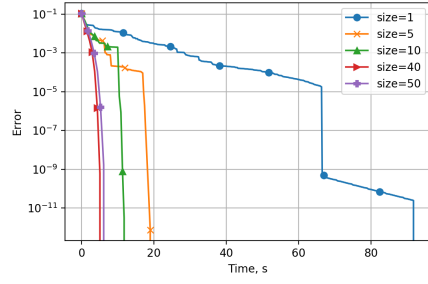
Figure 2: w8a, $\lambda = 10^{-3}$, $n = 49,749$, $d = 300$

Contrary to simulations in Section 4.1, we find that **svd** sketch is outperformed by others. Overall, **gauss** yielded the best results throughout all numerical tests.

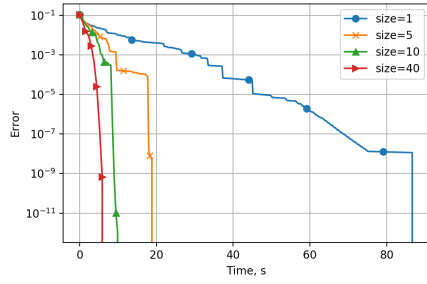
Randomized BFGS is comparable to the classical BFGS method on datasets with small d and moderate number



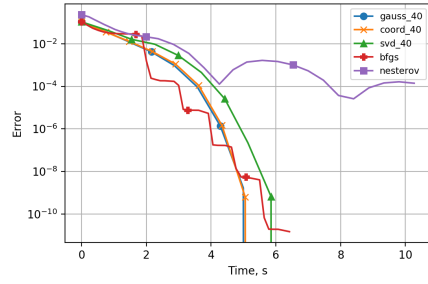
(a) gauss



(b) coord

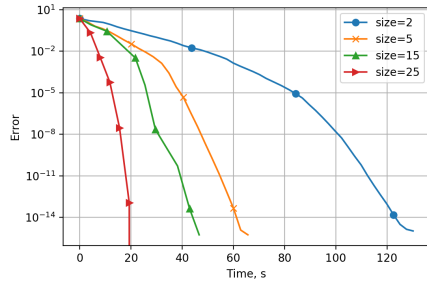


(c) svd

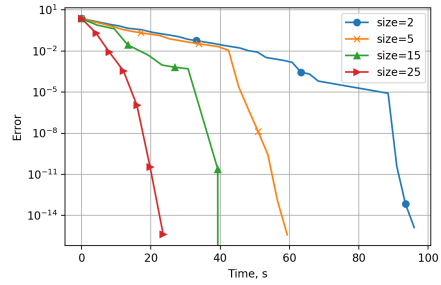


(d) methods compared

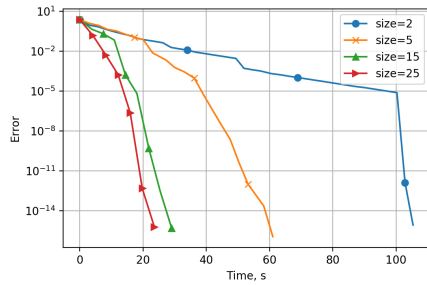
Figure 4: covtype, $\lambda = 10^{-3}$, $n = 581,012$, $d = 54$



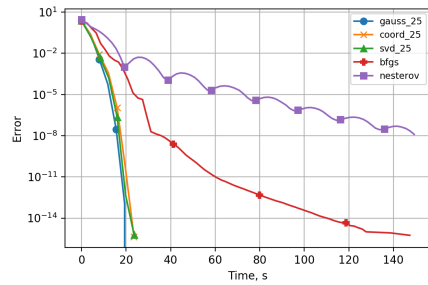
(a) gauss



(b) coord

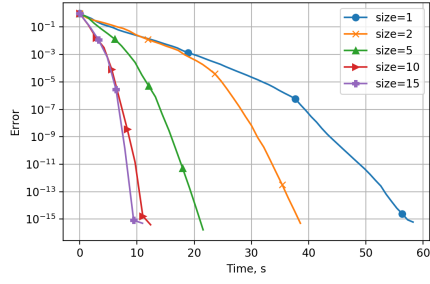


(c) svd

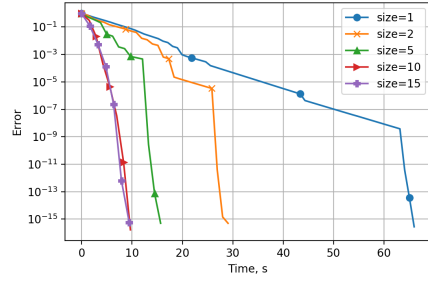


(d) methods compared

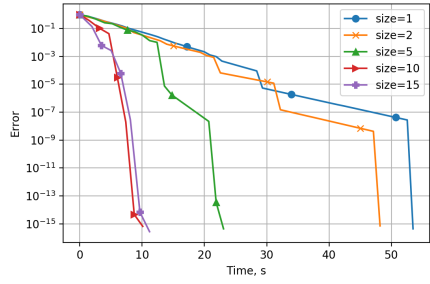
Figure 5: HIGGS, $\lambda = 10^{-1}$, $n = 11,000,000$, $d = 28$



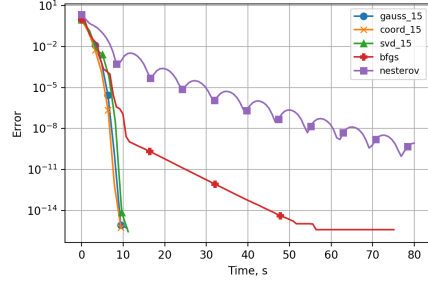
(a) gauss



(b) coord

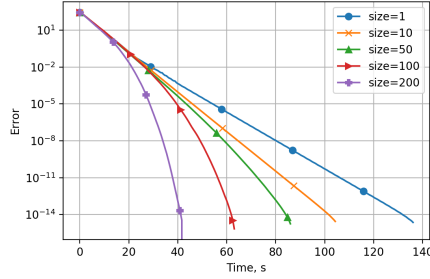


(c) svd

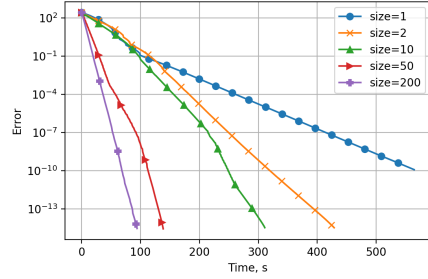


(d) methods compared

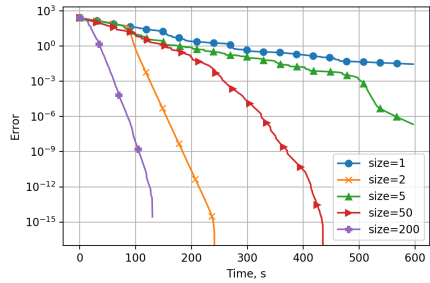
Figure 6: SUSY, $\lambda = 10^{-1}$, $n = 5,000,000$, $d = 18$



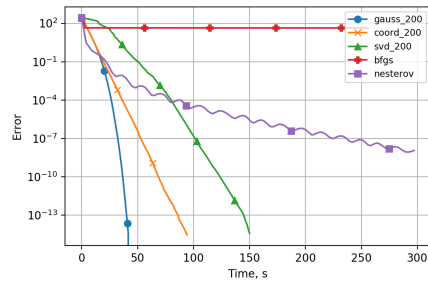
(a) gauss



(b) coord



(c) svd



(d) methods compared

Figure 7: gisette, $\lambda = 10^{-1}$, $n = 6,000$, $d = 5,000$

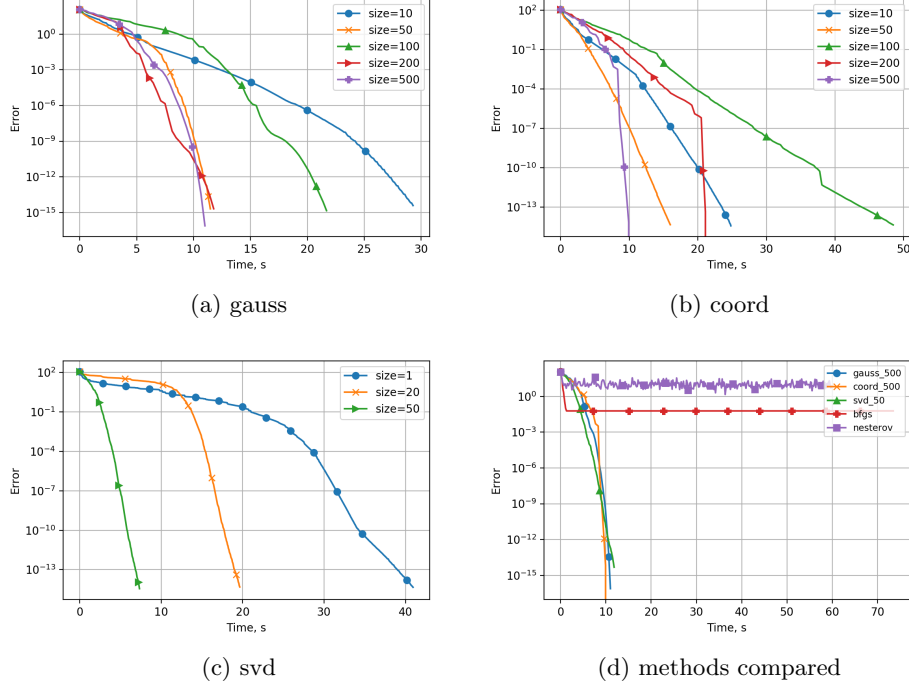


Figure 8: colon-cancer, $\lambda = 10^{-1}$, $n = 62$, $d = 2,000$

of samples, see Figures 2, 3, 4, but it performs better than classical BFGS on datasets with a high number of samples, see Figures 5, 6. Moreover, Randomized BFGS significantly outperforms the classical BFGS method on larger dimensional problems such as *gisette* ($d = 5000$), *colon-cancer* ($d = 2000$) and *epsilon* ($d = 2000$), see Figures 7, 8 and 9.

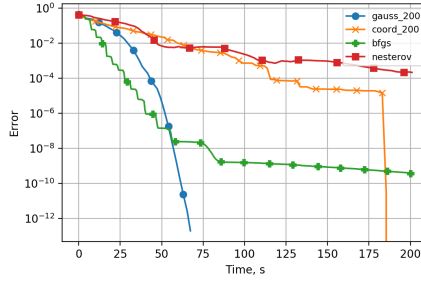


Figure 9: epsilon, $\lambda = 10^{-4}$, $n = 400,000$, $d = 2,000$

5 Conclusion, Consequences and Future Work

With a meaningful rate for the semi-local convergence of the randomized BFGS we now see several new open venues.

De-randomizing. One immediate question is whether randomization is necessary to establish a good convergence rate for the BFGS method. That is, can we show that the traditional deterministic BFGS method also enjoys a fast convergence rate? Based on our technique, we believe it may be possible through the following. First, we would need to establish a rate at which the estimates of the inverse matrix converge to a fixed constant Hessian matrix. In other words, we need to understand how fast would the BFGS matrices \mathbf{B}_k converge to the inverse Hessian were the BFGS method were applied to minimizing a quadratic. We believe this step could be answered by using

the equivalence of the BFGS method and the conjugate gradients method (CG) (Nazareth, 1979). Since the CG method has been shown to converge linearly at an accelerated rate (as compared to gradient descent), we suspect that the Hessian estimates of the BFGS method might also converge at an accelerated rate. If this is confirmed, the remaining steps of the proof would follow verbatim from our proofs.

Globalizing. By using a globalization strategy, such as a trust region (Conn et al., 2000), line-search or a carefully designed continuation scheme (Renegar, 2001), we believe it may be possible to extend our results to obtain a global convergence theory.

Practical consequences. The general rate of convergence and the bound in (30) suggest that new sketching matrices could be devised for accelerating our randomized BFGS method. We believe that by exploring the use of modern fast randomized sketches, such as the ROS sketch of Pilanci & Wainwright (2017), it may be possible to further accelerate the convergence of the randomized BFGS method.

Beyond self-concordance. It may be possible to extend our results beyond the class of self-concordant functions to the class of generalized self-concordant functions (Sun & Tran-Dinh, 2018; Bach, 2010) which also allow for one to control how fast the Hessian (and the inverse Hessian) changes.

References

- Bach, F. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(0):384–414, 2010.
- Broyden, C. G. Quasi-Newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.
- Broyden, C. G., Dennis, J. E., J., and Moré, J. J. On the Local and Superlinear Convergence of Quasi-Newton Methods. *IMA Journal of Applied Mathematics*, 12(3):223–245, 12 1973.
- Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.
- Chang, C.-C. and Lin, C.-J. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.
- Conn, A. R., Gould, N. I. M., and Toint, P. L. *Trust-region Methods*. Society for Industrial and Applied Mathematics, 2000.
- Fletcher, R. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–323, 1970.
- Floudas, C. C. A. and Pardalos, P. M. *Encyclopedia of Optimization*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387336249.
- Gao, W. and Goldfarb, D. Quasi-newton methods: Superlinear convergence without line search for self-concordant functions. *arXiv:1612.06965*, 2016.
- Goldfarb, D. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.
- Gower, R. M. and Richtárik, P. Linearly convergent randomized iterative methods for computing the pseudoinverse. *arXiv:1612.06255*, 2016.
- Gower, R. M. and Richtárik, P. Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1380–1409, 2017.
- Gower, R. M., Goldfarb, D., and Richtárik, P. Stochastic block BFGS: Squeezing more curvature out of data. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
- Gower, R. M., Kovalev, D., Lieder, F., and Richtárik, P. RSN: Randomized subspace newton. *NeurIPS*, 2019.

- Hilbert, D. Ein beitrage zur theorie des legendre'schen polynoms. In *Acta Mathematica*, volume 18, pp. 155–159, 1894. URL http://archive.ymsc.tsinghua.edu.cn/pacm_paperurl/20170108202957911035998.
- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pp. 315–323. Curran Associates, Inc., 2013.
- Meng, S., Vaswani, S., Laradji, I., Schmidt, M., and Lacoste-Julien, S. Fast and furious convergence: Stochastic second order methods under interpolation. *arXiv:1910.04920*, 2019.
- Moritz, P., Nishihara, R., and Jordan, M. I. A linearly-convergent stochastic L-BFGS algorithm. *arXiv:1508.02087v1*, 2015.
- Nazareth, L. A relationship between the bfgs and conjugate gradient algorithms and its implications for new algorithms. *SIAM Journal on Numerical Analysis*, 16(5):794–800, February 1979.
- Nesterov, Y. *Introductory Lectures on Convex Programming Volume I : Basic course*, volume I. Willey, 1998.
- Nesterov, Y. and Nemirovskii, A. *Interior Point Polynomial Algorithms in Convex Programming*. Studies in Applied Mathematics. Society for Industrial and Applied Mathematics, 1987.
- Pilanci, M. and Wainwright, M. J. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.
- Powell, M. J. D. On the Convergence of the Variable Metric Algorithm. *IMA Journal of Applied Mathematics*, 7(1): 21–36, 02 1971.
- Renegar, J. *A mathematical view of interior-point methods in convex optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. ISBN 0-89871-502-4.
- Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*, 24 (111):647–656, 1970.
- Sun, T. and Tran-Dinh, Q. Generalized self-concordant functions: A recipe for newton-type methods. *arXiv:1703.04599*, 2018.
- Wolfe, P. Convergence conditions for ascent methods. ii: some corrections. *SIAM Review*, 13(2):185–188, 1971.

Appendix

A Proof of Theorem 2.1

First we collect some more notation.

Further notation. For matrices $\mathbf{W} \in \mathbb{R}^{d \times d}$, let

$$\|\mathbf{W}\|_x := \sup_{v \neq 0} \frac{\|\mathbf{W}v\|_x}{\|v\|_x}$$

denote the induced norm. Note that $\|\mathbf{W}\|_x = \left\| \mathbf{H}_x^{1/2} \mathbf{W} \mathbf{H}_x^{-1/2} \right\|_2$ and thus $\|\cdot\|_x$ is sub-multiplicative. Indeed,

$$\|\mathbf{W}\mathbf{V}\|_x = \left\| \mathbf{H}_x^{1/2} \mathbf{W} \mathbf{V} \mathbf{H}_x^{-1/2} \right\|_2 = \left\| \mathbf{H}_x^{1/2} \mathbf{W} \mathbf{H}_x^{-1/2} \mathbf{H}_x^{1/2} \mathbf{V} \mathbf{H}_x^{-1/2} \right\|_2 \leq \|\mathbf{W}\|_x \|\mathbf{V}\|_x.$$

To further abbreviate our notation we will write $\|\mathbf{W}\|_* := \|\mathbf{W}\|_{x_*}$.

A.1 Properties of self-concordant functions

We now collect some known consequences of self-concordance and develop some additional properties that we need for our main proof.

Lemma A.1. For all $y \in \mathcal{B}_x^1$ we have that

$$\left\| \mathbf{H}_x^{-1} \mathbf{H}_y \right\|_x, \left\| \mathbf{H}_y^{-1} \mathbf{H}_x \right\|_x \leq \frac{1}{(1 - \|y - x\|_x)^2}, \quad (32)$$

and

$$\left\| \mathbf{I} - \mathbf{H}_x^{-1} \mathbf{H}_y \right\|_x, \left\| \mathbf{I} - \mathbf{H}_y^{-1} \mathbf{H}_x \right\|_x \leq \frac{1}{(1 - \|y - x\|_x)^2} - 1, \quad (33)$$

Proof. See Theorem 2.2.1 of Renegar (2001). □

Lemma A.2. If $\|x - x_*\|_x \leq 1$ then

$$\left\| x - x_* - \mathbf{H}_*^{-1} \nabla f(x) \right\|_* \leq \frac{\|x - x_*\|_*^2}{1 - \|x - x_*\|_*}, \quad (34)$$

and

$$\left\| \mathbf{H}_*^{-1} \nabla f(x_k) \right\|_* \leq \frac{\|x_* - x_k\|_*}{1 - \|x_k - x_*\|_*}. \quad (35)$$

Proof. Note that

$$\begin{aligned} x - x_* - \mathbf{H}_*^{-1} \nabla f(x) &= x - x_* - \int_0^1 \mathbf{H}_*^{-1} \mathbf{H}_{x_* + t(x-x_*)} (x - x_*) dt \\ &= \int_0^1 (\mathbf{I} - \mathbf{H}_*^{-1} \mathbf{H}_{x_* + t(x-x_*)}) (x - x_*) dt. \end{aligned} \quad (36)$$

Taking norms on both sides and using the sub-multiplicativity of the induced norm gives

$$\begin{aligned}
\|x - x_* - \mathbf{H}_*^{-1} \nabla f(x)\|_* &\stackrel{(36)}{=} \left\| \int_0^1 (\mathbf{I} - \mathbf{H}_*^{-1} \mathbf{H}_{x_*+t(x-x_*)}) (x - x_*) dt \right\|_* \\
&\leq \int_0^1 \|\mathbf{I} - \mathbf{H}_*^{-1} \mathbf{H}_{x_*+t(x-x_*)}\|_* \|x - x_*\|_* dt \\
&\stackrel{(33)}{\leq} \|x - x_*\|_* \int_0^1 \left(\frac{1}{(1-t\|x-x_*\|_{x_*})^2} - 1 \right) dt \\
&= \frac{\|x - x_*\|_*^2}{1 - \|x - x_*\|_{x_*}}.
\end{aligned}$$

The proof of (35) follows from (32) since

$$\|\mathbf{H}_*^{-1} \nabla f(x)\|_* \leq \|x_* - x\|_* \int_0^1 \|\mathbf{H}_*^{-1} \mathbf{H}_{x_*+t(x-x_*)}\|_* dt \stackrel{(32)}{\leq} \int_0^1 \frac{\|x_* - x\|_*}{(1-t\|x-x_*\|_*)^2} dt = \frac{\|x_* - x\|_*}{1 - \|x - x_*\|_*}.$$

□

Using self-concordance we can change the metric in the weight Frobenius.

Lemma A.3. Let $\mathbf{W} \in \mathbb{S}^d$. For all $y \in \mathcal{B}_x^1$ the following inequality holds:

$$\|\mathbf{W}\|_{F(\mathbf{H}_y)} \leq \frac{\|\mathbf{W}\|_{F(\mathbf{H}_x)}}{(1 - \|y - x\|_x)^2}. \tag{37}$$

Furthermore, if $x \in \mathcal{B}_y^{1/2}$ then

$$\|\mathbf{W}\|_{F(\mathbf{H}_y)} \leq \left(\frac{1 - \|y - x\|_y}{1 - 2\|y - x\|_y} \right)^2 \|\mathbf{W}\|_{F(\mathbf{H}_x)}. \tag{38}$$

Proof.

$$\begin{aligned}
\|\mathbf{W}\|_{F(\mathbf{H}_y)} &\leq \|\mathbf{W}\|_{F(\mathbf{H}_x)} \left\| \mathbf{H}_x^{-1/2} \mathbf{H}_y \mathbf{H}_x^{-1/2} \right\|_2 \\
&= \|\mathbf{W}\|_{F(\mathbf{H}_x)} \left\| \mathbf{H}_y \mathbf{H}_x^{-1} \right\|_x \\
&\stackrel{(32)}{\leq} \frac{\|\mathbf{W}\|_{F(\mathbf{H}_x)}}{(1 - \|y - x\|_x)^2}.
\end{aligned}$$

The bound (38) follows from the fact that for $x \in \mathcal{B}_y^{1/2}$ we have that $y \in \mathcal{B}_x^1$ and the definition of self-concordance (5) that

$$\|y - x\|_x \leq \frac{\|y - x\|_y}{1 - \|y - x\|_y} \quad \Rightarrow \quad \frac{1}{1 - \|y - x\|_x} \leq \frac{1 - \|y - x\|_y}{1 - 2\|y - x\|_y}.$$

□

A.2 The distance of the iterates

Next we need an upper bound on $\|x_{k+1} - x_*\|_*^2$.

Lemma A.4. If $x \in \mathcal{B}_{x_*}^1$ then

$$\|x_{k+1} - x_*\|_* \leq \frac{\|x_k - x_*\|_*^2}{2} \frac{3 - 2\|x_k - x_*\|_*}{(1 - \|x_k - x_*\|_*)^2} + \frac{1}{2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2. \tag{39}$$

Consequently if $x \in \mathcal{B}_{x_*}^{1/4}$ we have that

$$\|x_{k+1} - x_*\|_* \leq 5 \|x_k - x_*\|_*^2 + \frac{1}{2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}. \quad (40)$$

Proof. It starts with

$$\begin{aligned} \|x_{k+1} - x_*\|_* &= \|x_k - x_* - \mathbf{B}_k \nabla f(x_k)\|_* \\ &= \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k) + (\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_* \\ &\leq \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k)\|_* + \|(\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_* \\ &\leq \frac{\|x_k - x_*\|_*^2}{1 - \|x_k - x_*\|_*} + \|(\mathbf{B}_k - \mathbf{H}_*^{-1}) \nabla f(x_k)\|_*. \end{aligned}$$

where we used (34) in the last step. As for the second term we have that

$$\begin{aligned} \|(\mathbf{B}_k - \mathbf{H}_*^{-1}) \nabla f(x_k)\|_* &= \|(\mathbf{B}_k \mathbf{H}_* - \mathbf{I}) \mathbf{H}_*^{-1} \nabla f(x_k)\|_* \\ &\leq \|\mathbf{B}_k \mathbf{H}_* - \mathbf{I}\|_* \|\mathbf{H}_*^{-1} \nabla f(x_k)\|_* \\ &\stackrel{(35)}{\leq} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)} \frac{\|x_k - x_*\|_*}{1 - \|x_k - x_*\|_*}, \end{aligned}$$

where we used that

$$\|\mathbf{B}_k \mathbf{H}_* - \mathbf{I}\|_* = \|\mathbf{H}_*^{1/2} \mathbf{B}_k \mathbf{H}_*^{1/2} - \mathbf{I}\|_2 \leq \|\mathbf{H}_*^{1/2} \mathbf{B}_k \mathbf{H}_*^{1/2} - \mathbf{I}\|_F = \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}.$$

Finally using that $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ for all $a, b > 0$ with $a = \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}$ and $b = \frac{\|x_k - x_*\|_*}{1 - \|x_k - x_*\|_*}$ gives

$$\|(\mathbf{B}_k - \mathbf{H}_*^{-1}) \nabla f(x_k)\|_* \leq \frac{1}{2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 + \frac{1}{2} \frac{\|x_k - x_*\|_*^2}{(1 - \|x_k - x_*\|_*)^2}. \quad (41)$$

The above combined with (41) gives the result. \square

A.3 The distance of the quasi-Newton matrix

We start by establishing a lemma.

Lemma A.5. If $x \in \mathcal{B}_{x_*}^1$ then

$$\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + d \|x_k - x_*\|_*^2 \left(\frac{2 - \|x_k - x_*\|_*}{1 - \|x_k - x_*\|_*} \right)^2, \quad (42)$$

where $\rho := \sup_{x \in \mathbb{R}^d} \lambda_{\min} \left(\mathbb{E} \left[\mathbf{H}_x^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_x \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{H}_x^{1/2} \right] \right)$. Consequently, if $x \in \mathcal{B}_{x_*}^{1/4}$, then

$$\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + 3d \|x_k - x_*\|_*^2. \quad (43)$$

Proof. We define the following projection matrix:

$$\mathbf{Z}_k = \mathbf{H}_k^{1/2} \mathbf{S} (\mathbf{S}^\top \mathbf{H}_k \mathbf{S})^\dagger \mathbf{S}^\top \mathbf{H}_k^{1/2}. \quad (44)$$

Hence from (2) we have that

$$\mathbf{B}_{k+1} = \mathbf{H}_k^{-1/2} \left(\mathbf{Z}_k + (\mathbf{I} - \mathbf{Z}_k) \mathbf{H}_k^{1/2} \mathbf{B}_k \mathbf{H}_k^{1/2} (\mathbf{I} - \mathbf{Z}_k) \right) \mathbf{H}_k^{-1/2}. \quad (45)$$

To abbreviate the calculations we define the linear operator $P : \mathbb{S}^d \rightarrow \mathbb{S}^d$:

$$P(\mathbf{A}) := (\mathbf{I} - \mathbf{Z}_k) \mathbf{A} (\mathbf{I} - \mathbf{Z}_k). \quad (46)$$

Using this notation we rewrite $\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2$:

$$\begin{aligned}\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 &= \left\| \mathbf{H}_k^{1/2}(\mathbf{B}_{k+1} - \mathbf{H}_*^{-1})\mathbf{H}_k^{1/2} \right\|_F^2 \\ &= \left\| \mathbf{Z}_k + P(\mathbf{H}_k^{1/2}\mathbf{B}_k\mathbf{H}_k^{1/2}) - \mathbf{H}_k^{1/2}\mathbf{H}_*^{-1}\mathbf{H}_k^{1/2} \right\|_F^2 \\ &= \left\| P(\mathbf{H}_k^{1/2}(\mathbf{B}_k - \mathbf{H}_*^{-1})\mathbf{H}_k^{1/2}) + P(\mathbf{H}_k^{1/2}\mathbf{H}_*^{-1}\mathbf{H}_k^{1/2}) + \mathbf{Z}_k - \mathbf{H}_k^{1/2}\mathbf{H}_*^{-1}\mathbf{H}_k^{1/2} \right\|_F^2.\end{aligned}$$

Now note that the following equality holds

$$\mathbf{Z}_k = \mathbf{I} - P(\mathbf{I}) = \mathbf{H}_k^{1/2}\mathbf{H}_k^{-1}\mathbf{H}_k^{1/2} - P(\mathbf{H}_k^{1/2}\mathbf{H}_k^{-1}\mathbf{H}_k^{1/2}). \quad (47)$$

Using this equality and the shorthand

$$\mathbf{R}_k := \mathbf{H}_k^{1/2}(\mathbf{B}_k - \mathbf{H}_*^{-1})\mathbf{H}_k^{1/2} \in \mathbb{S}^d \quad \text{and} \quad \mathbf{D}_k := \mathbf{H}_k^{1/2}(\mathbf{H}_*^{-1} - \mathbf{H}_k^{-1})\mathbf{H}_k^{1/2} \in \mathbb{S}^d,$$

we have that

$$\begin{aligned}\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 &= \left\| P(\mathbf{H}_k^{1/2}(\mathbf{B}_k - \mathbf{H}_*^{-1})\mathbf{H}_k^{1/2}) + P(\mathbf{H}_k^{1/2}(\mathbf{H}_*^{-1} - \mathbf{H}_k^{-1})\mathbf{H}_k^{1/2}) - \mathbf{H}_k^{1/2}(\mathbf{H}_*^{-1} - \mathbf{H}_k^{-1})\mathbf{H}_k^{1/2} \right\|_F^2 \\ &= \underbrace{\|P(\mathbf{R}_k)\|_F^2}_{(I)} + \underbrace{\|P(\mathbf{D}_k) - \mathbf{D}_k\|_F^2}_{(II)} + \underbrace{2\langle P(\mathbf{R}_k), P(\mathbf{D}_k) - \mathbf{D}_k \rangle}_{(III)}.\end{aligned} \quad (48)$$

One can show that $(III) = 0$. Note that (I) has the following upper-bound:

$$\begin{aligned}\|P(\mathbf{R}_k)\|_F^2 &= \text{Trace}(\mathbf{R}_k(\mathbf{I} - \mathbf{Z}_k)\mathbf{R}_k(\mathbf{I} - \mathbf{Z}_k)) \\ &= \text{Trace}(\mathbf{R}_k(\mathbf{I} - \mathbf{Z}_k)\mathbf{R}_k) - \text{Trace}(\mathbf{Z}_k\mathbf{R}_k(\mathbf{I} - \mathbf{Z}_k)\mathbf{R}_k\mathbf{Z}_k) \\ &\leq \text{Trace}(\mathbf{R}_k(\mathbf{I} - \mathbf{Z}_k)\mathbf{R}_k).\end{aligned} \quad (49)$$

Furthermore, (II) has the following upper-bound:

$$\begin{aligned}\|P(\mathbf{D}_k) - \mathbf{D}_k\|_F^2 &= \text{Trace}\left([\mathbf{D}_k - (\mathbf{I} - \mathbf{Z}_k)\mathbf{D}_k(\mathbf{I} - \mathbf{Z}_k)]^2\right) \\ &= \|\mathbf{D}_k\|_F^2 - \text{Trace}((\mathbf{I} - \mathbf{Z}_k)\mathbf{D}_k(\mathbf{I} - \mathbf{Z}_k)\mathbf{D}_k(\mathbf{I} - \mathbf{Z}_k)) \\ &\leq \|\mathbf{D}_k\|_F^2 - \left\| \mathbf{H}_*^{-1/2}(\mathbf{H}_k - \mathbf{H}_*)\mathbf{H}_*^{-1/2} \right\|_F^2 \\ &= \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2.\end{aligned} \quad (50)$$

Using the above together with (49) and (48) gives

$$\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \|\mathbf{H}_k - \mathbf{H}_*\|_{F(\mathbf{H}_*^{-1})}^2. \quad (51)$$

Now we need to bound the term $\|\mathbf{H}_k - \mathbf{H}_*\|_{F(\mathbf{H}_*^{-1})}$ using the properties of self-concordant functions. For $x_k \in \mathcal{B}_*(x_*, 1)$, we have

$$\begin{aligned}\|\mathbf{H}_k - \mathbf{H}_*\|_{F(\mathbf{H}_*^{-1})} &= \left\| \mathbf{H}_*^{-1/2}(\mathbf{H}_k - \mathbf{H}_*)\mathbf{H}_*^{-1/2} \right\|_F^2 \\ &\leq d \left\| \mathbf{H}_*^{-1/2}(\mathbf{H}_k - \mathbf{H}_*)\mathbf{H}_*^{-1/2} \right\|_2^2 \\ &= d \left\| \mathbf{H}_*^{-1}\mathbf{H}_k - \mathbf{I} \right\|_*^2 \\ &\stackrel{(33)}{\leq} d \left(\frac{1}{(1 - \|x_k - x_*\|_*)^2} - 1 \right)^2 \\ &= d \|x_k - x_*\|_*^2 \left(\frac{2 - \|x_k - x_*\|_*}{1 - \|x_k - x_*\|_*} \right)^2.\end{aligned}$$

□

A.4 Detailed proof of Theorem 2.1

Proof. Taking expectation conditioned on x_k , using (38) and then Lemma A.5 gives

$$\begin{aligned}\mathbb{E}_k \left[\left\| \mathbf{B}_{k+1} - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_*)}^2 \right] &\leq \left(\frac{1 - \|x_k - x_*\|_*}{1 - 2\|x_k - x_*\|_*} \right)^2 \left\| \mathbf{B}_{k+1} - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_k)}^2 \\ &\leq \left(\frac{1 - \|x_k - x_*\|_*}{1 - 2\|x_k - x_*\|_*} \right)^2 \left((1 - \rho) \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_k)}^2 + 3d \|x_k - x_*\|_* \right) \\ &\leq \frac{1 - \rho}{(1 - 2\|x_k - x_*\|_*)^2} \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_k)}^2 + 3d \|x_k - x_*\|_* \left(\frac{1 - \|x_k - x_*\|_*}{1 - 2\|x_k - x_*\|_*} \right)^2.\end{aligned}$$

Now using that $\|x_k - x_*\|_* < \frac{1}{4}$ and combining the above with (40) gives

$$\Phi_\sigma^{k+1} \leq \sigma \left(\frac{1 - \rho}{(1 - 2\|x_k - x_*\|_*)^2} + \frac{1}{2\sigma} \right) \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_*)}^2 + (9\sigma d + 5) \|x_k - x_*\|_*^2.$$

Now note that

$$\|x_k - x_*\| \leq \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1 - \rho}{1 - \frac{2\rho}{3}}} \implies \frac{1 - \rho}{(1 - 2\|x_k - x_*\|_*)^2} \leq 1 - \frac{2\rho}{3},$$

and thus

$$\Phi_\sigma^{k+1} \leq \sigma \left(1 - \frac{2\rho}{3} + \frac{1}{2\sigma} \right) \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_*)}^2 + (9\sigma d + 5) \|x_k - x_*\|_*^2.$$

Now choosing $\sigma = \frac{3}{7\rho}$ and further restricting $\|x_k - x_*\|$ so that

$$\|x_k - x_*\|_* \leq \frac{2 - \rho}{9\rho d + 10} \implies (9\sigma d + 5) \|x_k - x_*\|_*^2 \leq \left(1 - \frac{\rho}{2} \right) \|x_k - x_*\|_*^2. \quad (52)$$

With this we have that

$$\mathbb{E} [\Phi_\sigma^{k+1}] \leq \sigma \left(1 - \frac{\rho}{2} \right) \left\| \mathbf{B}_k - \mathbf{H}_*^{-1} \right\|_{F(\mathbf{H}_*)}^2 + \left(1 - \frac{\rho}{2} \right) \|x_k - x_*\|_*^2 = \left(1 - \frac{\rho}{2} \right) \Phi_\sigma^k.$$

□

B Proof of Theorem 2.5

B.1 Getting ready for the proof

We first state and prove four lemmas.

Lemma B.1. Let $\mathbf{W} \in \mathbb{S}^d$. For all $x, y \in \mathbb{R}^d$ the following inequality holds:

$$\|\mathbf{W}\|_{F(\mathbf{H}_x)} \leq \left(1 + \frac{L_2}{\mu} \|x - y\|_2 \right) \|\mathbf{W}\|_{F(\mathbf{H}_y)}. \quad (53)$$

Proof.

$$\begin{aligned}\|\mathbf{W}\|_{F(\mathbf{H}_x)} &\leq \|\mathbf{W}\|_{F(\mathbf{H}_y)} \left\| \mathbf{H}_y^{-1/2} \mathbf{H}_x \mathbf{H}_y^{-1/2} \right\|_2 \\ &\leq \|\mathbf{W}\|_{F(\mathbf{H}_y)} \left(1 + \left\| \mathbf{H}_y^{-1/2} [\mathbf{H}_x - \mathbf{H}_y] \mathbf{H}_y^{-1/2} \right\|_2 \right) \\ &\leq \|\mathbf{W}\|_{F(\mathbf{H}_y)} \left(1 + \left\| \mathbf{H}_y^{-1} \right\|_2 \|\mathbf{H}_x - \mathbf{H}_y\|_2 \right) \\ &\leq \|\mathbf{W}\|_{F(\mathbf{H}_y)} (1 + \mu^{-1} L_2 \|x - y\|_2).\end{aligned}$$

□

Lemma B.2. Let $\mathbf{H}_* := \nabla^2 f(x_*)$. It follows that

$$\|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k)\|_2 \leq \frac{L_2}{2\mu} \|x_k - x_*\|_2^2.$$

Proof. By the fundamental theorem of calculus,

$$\nabla f(x_k) = \nabla f(x_k) - \nabla f(x_*) = \int_0^1 \nabla^2 f(x_* + t(x_k - x_*))(x_k - x_*) dt. \quad (54)$$

The result now follows from

$$\begin{aligned} \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k)\|_2 &\stackrel{(54)}{=} \left\| \mathbf{H}_*^{-1} \left(\mathbf{H}_* - \int_0^1 \nabla^2 f(x_* + t(x_k - x_*)) dt \right) (x_k - x_*) \right\|_2 \\ &\leq \|\mathbf{H}_*^{-1}\|_2 \left\| \mathbf{H}_* - \int_0^1 \nabla^2 f(x_* + t(x_k - x_*)) dt \right\|_2 \|x_k - x_*\|_2 \\ &= \|\mathbf{H}_*^{-1}\|_2 \left\| \int_0^1 [\mathbf{H}_* - \nabla^2 f(x_* + t(x_k - x_*))] dt \right\|_2 \|x_k - x_*\|_2 \\ &\leq \|\mathbf{H}_*^{-1}\|_2 \left[\int_0^1 \|\mathbf{H}_* - \nabla^2 f(x_* + t(x_k - x_*))\|_2 dt \right] \|x_k - x_*\|_2 \\ &\stackrel{(13)+(11)}{\leq} \frac{L_2}{\mu} \left(\int_0^1 \|t(x_k - x_*)\|_2 dt \right) \|x_k - x_*\|_2 \\ &= \frac{L_2}{2\mu} \|x_k - x_*\|_2^2. \end{aligned}$$

□

Lemma B.3.

$$\sqrt{f(x_+) - f(x_*)} \leq \frac{\sqrt{2L_1}L_2}{\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2. \quad (55)$$

Proof. We start with an upper-bound for $\|x_+ - x_*\|_2$:

$$\begin{aligned} \|x_+ - x_*\|_2 &= \|x_k - x_* - \mathbf{B}_k \nabla f(x_k)\|_2 \\ &= \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k) + (\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_2 \\ &\leq \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k)\|_2 + \|(\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_2 \\ &\stackrel{\text{Lemma B.2}+(12)}{\leq} \frac{L_2}{2\mu} \|x_k - x_*\|_2^2 + L_1 \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_2 \|x_k - x_*\|_2. \end{aligned}$$

Now using that $ab \leq \frac{c}{2}a^2 + \frac{1}{2c}b^2$ for every $a, b, c > 0$ with $a = \|x_k - x_*\|_2$, $b = \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_2$ and $c = \frac{L_2}{\mu L_1}$ in the above gives

$$\begin{aligned} \|x_+ - x_*\|_2 &\leq \frac{L_2}{\mu} \|x_k - x_*\|_2^2 + \frac{\mu L_1^2}{2L_2} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_2^2 \\ &\leq \frac{L_2}{\mu} \|x_k - x_*\|_2^2 + \frac{\mu L_1^2}{2L_2} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_k)}^2 \|\mathbf{H}_k^{-1}\|_2^2 \\ &\leq \frac{L_2}{\mu} \|x_k - x_*\|_2^2 + \frac{L_1^2}{2\mu L_2} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_k)}^2. \end{aligned}$$

Using smoothness (12), the above inequality and then strong convexity (11) in that order, we have that

$$\begin{aligned}
\sqrt{f(x_+) - f(x_*)} &\leq \sqrt{\frac{L_1}{2}} \|x_+ - x_*\|_2 \\
&\leq \sqrt{\frac{L_1}{2}} \frac{L_2}{\mu} \|x_k - x_*\|_2^2 + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_k)}^2 \\
&\leq \frac{\sqrt{2}L_1 L_2}{\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_k)}^2.
\end{aligned}$$

□

Lemma B.4.

$$\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \frac{2dL_2^2}{\mu^3} (f(x_k) - f(x_*)). \quad (56)$$

Proof. Recalling (51) we have that

$$\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \|\mathbf{H}_k - \mathbf{H}_*\|_{F(\mathbf{H}_*^{-1})}^2. \quad (57)$$

Now note that

$$\begin{aligned}
\left\| \mathbf{H}_*^{-1/2} (\mathbf{H}_k - \mathbf{H}_*) \mathbf{H}_*^{-1/2} \right\|_F^2 &\leq d \|\mathbf{H}_*^{-1}\|_2^2 \|\mathbf{H}_k - \mathbf{H}_*\|_2^2 \\
&\stackrel{(11)+(13)}{\leq} \frac{dL_2^2}{\mu^2} \|x_k - x_*\|_2^2.
\end{aligned} \quad (58)$$

Finally, taking expectation over (57) and using the above gives

$$\begin{aligned}
\mathbb{E}_k \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] &\leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \frac{dL_2^2}{\mu^2} \|x_k - x_*\|_2^2 \\
&\leq (1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \frac{2dL_2^2}{\mu^3} (f(x_k) - f(x_*)).
\end{aligned}$$

□

Now, we consider the following Lyapunov function:

$$\Psi_k = \sqrt{f(x_k) - f(x_*)} + \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2, \quad (59)$$

where $\beta = \frac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho}$.

B.2 The Proof

Having established several key lemmas, we are now ready to prove Theorem 2.5.

Proof.

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &= \sqrt{f(x_{k+1}) - f(x_*)} + \beta \mathbb{E} \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 \right] \\
&\leq \sqrt{f(x_k) - f(x_*)} + \beta \mathbb{E} \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 \right] \\
&\stackrel{(55)+(53)}{\leq} \frac{\sqrt{2L_1}L_2}{\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \\
&\quad + \beta \left[1 + \frac{L_2}{\mu} \|x_k - x_*\|_2 \right]^2 \mathbb{E} \left[\|\mathbf{B}_{k+1} - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \right] \\
&\stackrel{(56)}{\leq} \frac{\sqrt{2L_1}L_2}{\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{5/2}}{2\sqrt{2}\mu L_2} \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 \\
&\quad + \beta \left[1 + \frac{L_2}{\mu} \|x_k - x_*\|_2 \right]^2 \left[(1 - \rho) \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_k)}^2 + \frac{2dL_2^2}{\mu^3} (f(x_k) - f(x_*)) \right] \\
&\leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{2\beta dL_2^2}{\mu^3} \left[1 + \frac{L_2}{\mu} \|x_k - x_*\|_2 \right]^2 \right] (f(x_k) - f(x_*)) \\
&\quad + \left[(1 - \rho) \left[1 + \frac{L_2}{\mu} \|x_k - x_*\|_2 \right]^4 + \frac{L_1^{5/2}}{2\beta\sqrt{2}\mu L_2} \left[1 + \frac{L_2}{\mu} \|x_k - x_*\|_2 \right]^2 \right] \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2.
\end{aligned}$$

Now, since $f(x_k) - f(x_*)$ is non-increasing by line 1 in Algorithm 1, we have that $\|x_k - x_*\|_2 \leq \sqrt{\frac{2}{\mu}(f(x_k) - f(x_*))} \leq \sqrt{\frac{2}{\mu}(f(x_0) - f(x_*))} \leq \sqrt{\frac{2\mathcal{F}}{\mu}}$:

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &\leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{2\beta dL_2^2}{\mu^3} \left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \right]^2 \right] (f(x_k) - f(x_*)) \\
&\quad + \left[(1 - \rho) \left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \right]^4 + \frac{L_1^{5/2}}{2\beta\sqrt{2}\mu L_2} \left[1 + \frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \right]^2 \right] \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2.
\end{aligned}$$

Equation (14) implies $\frac{\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \leq 1$ and we get

$$\begin{aligned}
\mathbb{E}[\Psi_{k+1}] &\leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{8\beta dL_2^2}{\mu^3} \right] (f(x_k) - f(x_*)) + \left[(1 - \rho) \left[1 + \frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \right] + \frac{\sqrt{2}L_1^{5/2}}{\beta\mu L_2} \right] \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 \\
&\leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{8\beta dL_2^2}{\mu^3} \right] (f(x_k) - f(x_*)) + \left[1 - \rho + \frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} + \frac{\sqrt{2}L_1^{5/2}}{\beta\mu L_2} \right] \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2.
\end{aligned}$$

Using $\beta = \frac{4\sqrt{2}L_1^{5/2}}{\mu L_2 \rho}$ and using $\frac{15\sqrt{2\mathcal{F}}L_2}{\mu^{3/2}} \leq \frac{\rho}{4}$, which follows from (14), we get

$$\mathbb{E}[\Psi_{k+1}] \leq \left[\frac{\sqrt{2L_1}L_2}{\mu^2} + \frac{32\sqrt{2}dL_1^{5/2}L_2}{\rho\mu^4} \right] \sqrt{\mathcal{F} \cdot (f(x_k) - f(x_*))} + \left(1 - \frac{\rho}{2} \right) \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2.$$

Finally, using (14) we get

$$\mathbb{E}[\Psi_{k+1}] \leq \frac{1}{2} \sqrt{(f(x_k) - f(x_*))} + \left(1 - \frac{\rho}{2} \right) \beta \|\mathbf{B}_k - \mathbf{H}_*^{-1}\|_{F(\mathbf{H}_*)}^2 \leq \left(1 - \frac{\rho}{2} \right) \Psi_k. \quad (60)$$

□

C Proof of Theorem 2.6

Lemma C.1. $\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}}$ converges to 0 with probability 1.

Proof. We start with an upper-bound for $\sqrt{f(x_{k+1})-f(x_*)}$:

$$\begin{aligned}
\sqrt{f(x_{k+1})-f(x_*)} &\leq \sqrt{f(x_+) - f(x_*)} \\
&\leq \sqrt{\frac{L_1}{2}} \|x_+ - x_*\|_2 \\
&= \sqrt{\frac{L_1}{2}} \|x_k - x_* - \mathbf{B}_k \nabla f(x_k)\|_2 \\
&= \sqrt{\frac{L_1}{2}} \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k) + (\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_2 \\
&\leq \sqrt{\frac{L_1}{2}} \|x_k - x_* - \mathbf{H}_*^{-1} \nabla f(x_k)\|_2 + \sqrt{\frac{L_1}{2}} \|(\mathbf{H}_*^{-1} - \mathbf{B}_k) \nabla f(x_k)\|_2 \\
&\stackrel{\text{Lemma B.2+(12)}}{\leq} \frac{L_2 \sqrt{L_1}}{2\sqrt{2}\mu} \|x_k - x_*\|_2^2 + \frac{L_1^{3/2}}{\sqrt{2}} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_2 \|x_k - x_*\|_2 \\
&\stackrel{(11)}{\leq} \frac{L_2 \sqrt{L_1}}{\sqrt{2}\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{3/2}}{\sqrt{\mu}} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_2 \sqrt{f(x_k) - f(x_*)} \\
&\leq \frac{L_2 \sqrt{L_1}}{\sqrt{2}\mu^2} (f(x_k) - f(x_*)) + \frac{L_1^{3/2}}{\mu^{3/2}} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_*)} \sqrt{f(x_k) - f(x_*)}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}} &\leq \frac{L_2 \sqrt{L_1}}{\sqrt{2}\mu^2} \sqrt{f(x_k) - f(x_*)} + \frac{L_1^{3/2}}{\mu^{3/2}} \|\mathbf{H}_*^{-1} - \mathbf{B}_k\|_{F(\mathbf{H}_*)} \\
&\leq a\Psi_k + b\sqrt{\Psi_k},
\end{aligned}$$

where $a = \frac{L_2 \sqrt{L_1}}{\sqrt{2}\mu^2}$ and $b = \frac{L_1^{3/2}}{\mu^{3/2}\sqrt{\beta}}$. After taking expectation, we get

$$\begin{aligned}
\mathbb{E} \left[\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}} \right] &\leq \mathbb{E} [a\Psi_k + b\sqrt{\Psi_k}] \\
&\leq a\mathbb{E} [\Psi_k] + b\sqrt{\mathbb{E} [\Psi_k]} \\
&\leq a(1-\rho)^k \Psi_0 + b\sqrt{(1-\rho)^k \Psi_0} \\
&\leq cq^k,
\end{aligned}$$

where $q = \sqrt{1-\rho} \in (0, 1)$ and $c = a\Psi_0 + b\sqrt{\Psi_0}$. Now, we choose arbitrary ϵ and apply Markov's inequality:

$$\mathbb{P} \left(\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}} \geq \epsilon \right) \leq \frac{1}{\epsilon} \mathbb{E} \left[\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}} \right] \leq \frac{cq^k}{\epsilon}.$$

Hence, for all $\epsilon > 0$

$$\sum_{k=0}^{\infty} \mathbb{P} \left(\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}} \geq \epsilon \right) < \infty, \tag{61}$$

which means that $\sqrt{\frac{f(x_{k+1})-f(x_*)}{f(x_k)-f(x_*)}}$ converges to 0 in probability sufficiently quickly and hence it converges to 0 almost surely. \square

Lemma C.2. $\sqrt{f(x_k) - f(x_*)}$ converges to 0 with probability 1.

Proof. $\mathbb{E} \left[\sqrt{f(x_k) - f(x_*)} \right]$ converges linearly to 0, thus it converges in probability to 0 sufficiently quickly. Hence, $\sqrt{f(x_k) - f(x_*)}$ converges to 0 almost surely. \square