

Newton Sketch: A Linear-time Optimization Algorithm with Linear-Quadratic Convergence

Mert Pilanci¹ Martin J. Wainwright^{1,2}

{mert, wainwrig}@berkeley.edu

University of California, Berkeley

¹Department of Electrical Engineering and Computer Science

²Department of Statistics

May 12, 2015

Abstract

We propose a randomized second-order method for optimization known as the Newton Sketch: it is based on performing an approximate Newton step using a randomly projected or sub-sampled Hessian. For self-concordant functions, we prove that the algorithm has super-linear convergence with exponentially high probability, with convergence and complexity guarantees that are independent of condition numbers and related problem-dependent quantities. Given a suitable initialization, similar guarantees also hold for strongly convex and smooth objectives without self-concordance. When implemented using randomized projections based on a sub-sampled Hadamard basis, the algorithm typically has substantially lower complexity than Newton’s method. We also describe extensions of our methods to programs involving convex constraints that are equipped with self-concordant barriers. We discuss and illustrate applications to linear programs, quadratic programs with convex constraints, logistic regression and other generalized linear models, as well as semidefinite programs.

1 Introduction

Relative to first-order methods, second-order methods for convex optimization enjoy superior convergence in both theory and practice. For instance, Newton’s method converges at a quadratic rate for strongly convex and smooth problems, and moreover, even for weakly convex functions (i.e. not strongly convex), modifications of Newton’s method has super-linear convergence compared to the much slower $1/T^2$ convergence rate that can be achieved by a first-order method like accelerated gradient descent (see e.g. [15]). More importantly, at least in a uniform sense, the $1/T^2$ -rate is known to be unimprovable for first-order methods [17]. Yet another issue in first-order methods is the tuning of step size, whose optimal choice depends on the strong convexity parameter and/or smoothness of the underlying problem. For example, consider the problem of optimizing a function of the form $x \mapsto g(Ax)$, where $A \in \mathbb{R}^{n \times d}$ is a “data matrix”, and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a twice-differentiable function. Here the performance of first-order methods will depend on both the convexity/smoothness of g , as well as the conditioning of the data matrix. In contrast, whenever the function g is self-concordant, then Newton’s method with suitably damped steps has a global complexity guarantee that is provably independent of such problem-dependent parameters.

On the other hand, each step of Newton’s method requires solving a linear system defined by the Hessian matrix. For instance, in application to the problem family just described involving an $n \times d$ data matrix, each of these steps has complexity scaling as $\mathcal{O}(nd^2)$. For this reason, both forming the Hessian and solving the corresponding linear system pose a tremendous numerical challenge for large values of (n, d) — for instance, values of thousands to millions, as is common in big data applications. In order to address this issue, a multitude of different approximations to Newton’s method have been proposed and studied in the literature. Quasi-Newton methods form estimates of the Hessian by successive evaluations of the gradient vectors and are computationally cheaper. Examples of such methods include DFP and BFGS schemes and also their limited memory versions (see the book [25] for further details). A disadvantage of such approximations based on first-order information is that the associated convergence guarantees are typically much weaker than those of Newton’s method and require stronger assumptions. Under restrictions on the eigenvalues of the Hessian (strong convexity and smoothness), Quasi-Newton methods typically exhibit local super-linear convergence.

In this paper, we propose and analyze a randomized approximation of Newton’s method, known as the *Newton Sketch*. Instead of explicitly computing the Hessian, the Newton Sketch method approximates it via a random projection of dimension m . When these projections are carried out using the randomized Hadamard transform, each iteration has complexity $\mathcal{O}(nd \log(m) + dm^2)$. Our results show that it is always sufficient to choose m proportional to $\min\{d, n\}$, and moreover, that the sketch dimension m can be much smaller for certain types of constrained problems. Thus, in the regime $n > d$ and with $m \asymp d$, the complexity per iteration can be substantially lower than the $\mathcal{O}(nd^2)$ complexity of each Newton step. Specifically for $n \geq d^2$, the complexity of Newton Sketch per iteration is $\mathcal{O}(nd \log d)$, which is linear in the input size (nd) and comparable to first order methods which only access the derivative $g'(Ax)$. Moreover, we show that for self-concordant functions, the total complexity of obtaining a δ -optimal solution is $\mathcal{O}(nd \log d \log(1/\delta))$, and does not depend on constants such as strong convexity or smoothness parameters unlike first order methods. On the other hand, for problems with $d > n$, we also provide a dual strategy which effectively has the same guarantees with roles of d and n exchanged.

We also consider other random projection matrices and sub-sampling strategies, including partial forms of random projection that exploit known structure in the Hessian. For self-concordant functions, we provide an affine invariant analysis proving that the convergence is linear-quadratic and the guarantees are independent of the function and data, such as condition numbers of matrices involved in the objective function. Finally, we describe an interior point method to deal with arbitrary convex constraints which combines the Newton sketch with the barrier method. We provide an upper bound on the total number of iterations required to obtain a solution with a pre-specified target accuracy.

The remainder of this paper is organized as follows. We begin in Section 2 with some background on the classical form of Newton’s method, random matrices for sketching, and Gaussian widths as a measure of the size of a set. In Section 3, we formally introduce the Newton Sketch, including both fully and partially sketched versions for unconstrained and constrained problems. We provide some illustrative examples in Section 3.2 before turning to local convergence theory in Section 3.3. Section 4 is devoted to global convergence results for self-concordant functions, in both the constrained and unconstrained settings. In Section 5, we consider a number of applications and provide additional numerical results. The bulk of our proofs are given in Section 6, with some more technical aspects deferred to the appendices.

2 Background

We begin with some background material on the standard form of Newton’s method, various types of random sketches, and the notion of Gaussian width as a complexity measure.

2.1 Classical version of Newton’s method

In this section, we briefly review the convergence properties and complexity of the classical form of Newton’s method; see the sources [25, 4, 17] for further background.

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a closed, convex and twice-differentiable function that is bounded below. Given a convex set \mathcal{C} , we assume that the constrained minimizer

$$x^* := \arg \min_{x \in \mathcal{C}} f(x) \quad (1)$$

is uniquely defined, and we define the minimum and maximum eigenvalues $\gamma = \lambda_{\min}(\nabla^2 f(x^*))$ and $\beta = \lambda_{\max}(\nabla^2 f(x^*))$ of the Hessian evaluated at the minimum.

We assume moreover that the Hessian map $x \mapsto \nabla^2 f(x)$ is Lipschitz continuous with modulus L , meaning that

$$\|\nabla^2 f(x + \Delta) - \nabla^2 f(x)\|_{\text{op}} \leq L \|\Delta\|_2. \quad (2)$$

Under these conditions and given an initial point $\tilde{x}^0 \in \mathcal{C}$ such that $\|\tilde{x}^0 - x^*\|_2 \leq \frac{\gamma}{2L}$, the Newton updates are guaranteed to converge quadratically—viz.

$$\|\tilde{x}^{t+1} - x^*\|_2 \leq \frac{2L}{\gamma} \|\tilde{x}^t - x^*\|_2^2,$$

This result is classical: for instance, see Boyd and Vandenberghe [4] for a proof. Newton’s method can be slightly modified to be globally convergent by choosing the step sizes via a simple backtracking line-search procedure.

The following result characterizes the complexity of Newton’s method when applied to self-concordant functions and is central in the development of interior point methods (for instance, see the books [18, 4]). We defer the definitions of self-concordance and the line-search procedure in the following sections. The number of iterations needed to obtain a δ approximate minimizer of a strictly convex self-concordant function f is bounded by

$$\frac{20 - 8a}{ab(1 - 2a)} (f(x^0) - f(x^*)) + \log_2 \log_2(1/\delta),$$

where a, b are constants in the line-search procedure.¹

2.2 Different types of randomized sketches

Various types of randomized sketches are possible, and we describe a few of them here. Given a sketching matrix $S \in \mathbb{R}^{m \times n}$, we use $\{s_i\}_{i=1}^m$ to denote the collection of its n -dimensional rows. We restrict our attention to sketch matrices that are zero-mean, and that are normalized so that $\mathbb{E}[S^T S/m] = I_n$.

¹Typical values of these constants are $a = 0.1$ and $b = 0.5$.

Sub-Gaussian sketches: The most classical sketch is based on a random matrix $S \in \mathbb{R}^{m \times n}$ with i.i.d. standard Gaussian entries, or somewhat more generally, sketch matrices based on i.i.d. sub-Gaussian rows. In particular, a zero-mean random vector $s \in \mathbb{R}^n$ is 1-sub-Gaussian if for any $u \in \mathbb{R}^n$, we have

$$\mathbb{P}[\langle s, u \rangle \geq \epsilon \|u\|_2] \leq e^{-\epsilon^2/2} \quad \text{for all } \epsilon \geq 0. \quad (3)$$

For instance, a vector with i.i.d. $N(0, 1)$ entries is 1-sub-Gaussian, as is a vector with i.i.d. Rademacher entries (uniformly distributed over $\{-1, +1\}$). We use the terminology *sub-Gaussian sketch* to mean a random matrix $S \in \mathbb{R}^{m \times n}$ with i.i.d. rows that are zero-mean, 1-sub-Gaussian, and with $\text{cov}(s) = I_n$.

From a theoretical perspective, sub-Gaussian sketches are attractive because of the well-known concentration properties of sub-Gaussian random matrices (e.g., [5, 24]). On the other hand, from a computational perspective, a disadvantage of sub-Gaussian sketches is that they require matrix-vector multiplications with unstructured random matrices. In particular, given a data matrix $A \in \mathbb{R}^{n \times d}$, computing its sketched version SA requires $\mathcal{O}(mnd)$ basic operations in general (using classical matrix multiplication).

Sketches based on randomized orthonormal systems (ROS): The second type of randomized sketch we consider is *randomized orthonormal system* (ROS), for which matrix multiplication can be performed much more efficiently. In order to define a ROS sketch, we first let $H \in \mathbb{R}^{n \times n}$ be an orthonormal matrix with entries $H_{ij} \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$. Standard classes of such matrices are the Hadamard or Fourier bases, for which matrix-vector multiplication can be performed in $\mathcal{O}(n \log n)$ time via the fast Hadamard or Fourier transforms, respectively. Based on any such matrix, a sketching matrix $S \in \mathbb{R}^{m \times n}$ from a ROS ensemble is obtained by sampling i.i.d. rows of the form

$$s^T = \sqrt{n} e_j^T H D \quad \text{with probability } 1/n \text{ for } j = 1, \dots, n,$$

where the random vector $e_j \in \mathbb{R}^n$ is chosen uniformly at random from the set of all n canonical basis vectors, and $D = \text{diag}(\nu)$ is a diagonal matrix of i.i.d. Rademacher variables $\nu \in \{-1, +1\}^n$. Given a fast routine for matrix-vector multiplication, the sketch SM for a data matrix $M \in \mathbb{R}^{n \times d}$ can be formed in $\mathcal{O}(nd \log m)$ time (for instance, see the papers [2, 1]).

Sketches based on random row sampling: Given a probability distribution $\{p_j\}_{j=1}^n$ over $[n] = \{1, \dots, n\}$, another choice of sketch is to randomly sample the rows of a data matrix M a total of m times with replacement from the given probability distribution. Thus, the rows of S are independent and take on the values

$$s^T = \frac{e_j}{\sqrt{p_j}} \quad \text{with probability } p_j \text{ for } j = 1, \dots, n$$

where $e_j \in \mathbb{R}^n$ is the j^{th} canonical basis vector. Different choices of the weights $\{p_j\}_{j=1}^n$ are possible, including those based on the row ℓ_2 norms $p_j \propto \|Me_j\|_2^2$ and leverage values of M —i.e., $p_j \propto \|Ue_j\|_2$ for $j = 1, \dots, n$, where $U \in \mathbb{R}^{n \times d}$ is the matrix of left singular vectors of M [6]. When $M \in \mathbb{R}^{n \times d}$ is the adjacency matrix of a graph with d vertices and n edges, the leverage scores of M are also known as effective resistances which can be used to sub-sample edges of a given graph by preserving its spectral properties [22].

2.3 Gaussian widths

In this section, we introduce some background on the notion of Gaussian width, a way of measuring the size of a compact set in \mathbb{R}^d . These width measures play a key role in the analysis of randomized sketches. Given a compact subset $\mathcal{L} \subseteq \mathbb{R}^d$, its Gaussian width is given by

$$\mathcal{W}(\mathcal{L}) := \mathbb{E}_g \left[\max_{z \in \mathcal{L}} |\langle g, z \rangle| \right] \quad (4)$$

where $g \in \mathbb{R}^n$ is an i.i.d. sequence of $N(0, 1)$ variables. This complexity measure plays an important role in Banach space theory, learning theory and statistics (e.g., [21, 12, 3]).

Of particular interest in this paper are sets \mathcal{L} that are obtained by intersecting a given cone \mathcal{K} with the Euclidean sphere $\mathcal{S}^{d-1} = \{z \in \mathbb{R}^n \mid \|z\|_2 = 1\}$. It is easy to show that the Gaussian width of any such set is at most \sqrt{d} , but it can be substantially smaller, depending on the nature of the underlying cone. For instance, if \mathcal{K} is a subspace of dimension $r < d$, then a simple calculation yields that $\mathcal{W}(\mathcal{K} \cap \mathcal{S}^{d-1}) \leq \sqrt{r}$.

3 Newton sketch and local convergence

With the basic background in place, let us now introduce the Newton sketch algorithm, and then develop a number of convergence guarantees associated with it. **It applies to an optimization problem of the form $\min_{x \in \mathcal{C}} f(x)$, where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a twice-differentiable convex function, and $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex constraint set.**

3.1 Newton sketch algorithm

In order to motivate the Newton sketch algorithm, recall the standard form of Newton's algorithm: given a current iterate $\tilde{x}^t \in \mathcal{C}$, it generates the new iterate \tilde{x}^{t+1} by performing a constrained minimization of the second order Taylor expansion—viz.

$$\tilde{x}^{t+1} = \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2} \langle x - \tilde{x}^t, \nabla^2 f(\tilde{x}^t) (x - \tilde{x}^t) \rangle + \langle \nabla f(\tilde{x}^t), x - \tilde{x}^t \rangle \right\}. \quad (5a)$$

In the unconstrained case—that is, when $\mathcal{C} = \mathbb{R}^d$ —it takes the simpler form

$$\tilde{x}^{t+1} = \tilde{x}^t - [\nabla^2 f(\tilde{x}^t)]^{-1} \nabla f(\tilde{x}^t). \quad (5b)$$

Now suppose that we have available a Hessian matrix square root $\nabla^2 f(x)^{1/2}$ —that is, a matrix $\nabla^2 f(x)^{1/2}$ of dimensions $n \times d$ such that

$$(\nabla^2 f(x)^{1/2})^T \nabla^2 f(x)^{1/2} = \nabla^2 f(x) \quad \text{for some integer } n \geq \text{rank}(\nabla^2 f(x)).$$

In many cases, such a matrix square root can be computed efficiently. For instance, consider a function of the form $f(x) = g(Ax)$ where $A \in \mathbb{R}^{n \times d}$, and the function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ has the separable form $g(Ax) = \sum_{i=1}^n g_i(\langle a_i, x \rangle)$. In this case, a suitable Hessian matrix square root is given by the $n \times d$ matrix $\nabla^2 f(x)^{1/2} := \text{diag}\{g_i''(\langle a_i, x \rangle)\}_{i=1}^n A$. In Section 3.2, we discuss various concrete instantiations of such functions.

In terms of this notation, the ordinary Newton update can be re-written as

$$\tilde{x}^{t+1} = \arg \min_{x \in \mathcal{C}} \underbrace{\left\{ \frac{1}{2} \|\nabla^2 f(\tilde{x}^t)^{1/2} (x - \tilde{x}^t)\|_2^2 + \langle \nabla f(\tilde{x}^t), x - \tilde{x}^t \rangle \right\}}_{\tilde{\Phi}(x)},$$

and the Newton Sketch algorithm is most easily understood based on this form of the updates. More precisely, for a sketch dimension m to be chosen, let $S \in \mathbb{R}^{m \times n}$ be an isotropic sketch matrix, satisfying the relation $\mathbb{E}[S^T S] = I_n$. The *Newton Sketch algorithm* generates a sequence of iterates $\{x^t\}_{t=0}^\infty$ according to the recursion

$$x^{t+1} := \arg \min_{x \in \mathcal{C}} \underbrace{\left\{ \frac{1}{2} \|S^t \nabla^2 f(x^t)^{1/2} (x - x^t)\|_2^2 + \langle \nabla f(x^t), x - x^t \rangle \right\}}_{\Phi(x; S^t)}, \quad (6)$$

where $S^t \in \mathbb{R}^{m \times d}$ is an independent realization of a sketching matrix. When the problem is unconstrained, i.e., $\mathcal{C} = \mathbb{R}^d$ and the matrix $\nabla^2 f(x^t)^{1/2} (S^t)^T S^t \nabla^2 f(x^t)^{1/2}$ is invertible, the Newton sketch update takes the simpler form to

$$x^{t+1} = x^t - \left(\nabla^2 f(x^t)^{1/2} (S^t)^T S^t \nabla^2 f(x^t)^{1/2} \right)^{-1} \nabla f(x^t). \quad (7)$$

The intuition underlying the Newton sketch updates is as follows: the iterate x^{t+1} corresponds to the constrained minimizer of the random objective function $\Phi(x; S^t)$ whose expectation $\mathbb{E}[\Phi(x; S^t)]$, taking averages over the isotropic sketch matrix S^t , is equal to the original Newton objective $\tilde{\Phi}(x)$. Consequently, it can be seen as a stochastic form of the Newton update.

In this paper, we also analyze a *partially sketched Newton update*, which takes the following form. Given an additive decomposition of the form $f = f_0 + g$, we perform a sketch of of the Hessian $\nabla^2 f_0$ while retaining the exact form of the Hessian $\nabla^2 g$. This leads to the partially sketched update

$$x^{t+1} := \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2} (x - x^t)^T Q^t (x - x^t) + \langle \nabla f(x^t), x - x^t \rangle \right\} \quad (8)$$

where $Q^t := (S^t \nabla^2 f_0(x^t)^{1/2})^T S^t \nabla^2 f_0(x^t)^{1/2} + \nabla^2 g(x^t)$.

For either the fully sketched (6) or partially sketched updates (8), our analysis shows that there are many settings in which the sketch dimension m can be chosen to be substantially smaller than n , in which cases the sketched Newton updates will be much cheaper than a standard Newton update. For instance, the unconstrained update (7) can be computed in at most $\mathcal{O}(md^2)$ time, as opposed to the $\mathcal{O}(nd^2)$ time of the standard Newton update. In constrained settings, we show that the sketch dimension m can often be chosen even smaller—even $m \ll d$ —which leads to further savings.

3.2 Some examples

In order to provide some intuition, let us provide some simple examples to which the sketched Newton updates can be applied.

Example 1 (Newton sketch for LP solving). Consider a linear program (LP) in the standard form

$$\min_{Ax \leq b} \langle c, x \rangle \quad (9)$$

where $A \in \mathbb{R}^{n \times d}$ is a given constraint matrix. We assume that the polytope $\{x \in \mathbb{R}^d \mid Ax \leq b\}$ is bounded so that the minimum is achieved. A barrier method approach to this LP is based on

solving a sequence of problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ \underbrace{\tau \langle c, x \rangle - \sum_{i=1}^n \log(b_i - \langle a_i, x \rangle)}_{f(x)} \right\},$$

where $a_i \in \mathbb{R}^d$ denotes the i^{th} row of A , and $\tau > 0$ is a weight parameter that is adjusted during the algorithm. By inspection, the function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is twice-differentiable, and its Hessian is given by $\nabla^2 f(x) = A^T \text{diag}\left\{\frac{1}{(b_i - \langle a_i, x \rangle)^2}\right\} A$. A Hessian square root is given by $\nabla^2 f(x)^{1/2} := \text{diag}\left(\frac{1}{|b_i - \langle a_i, x \rangle|}\right) A$, which allows us to compute a sketched version of the Hessian square root

$$S \nabla^2 f(x)^{1/2} = S \text{diag}\left(\frac{1}{|b_i - \langle a_i, x \rangle|}\right) A.$$

With a ROS sketch matrix, computing this matrix requires $\mathcal{O}(nd \log(m))$ basic operations. The complexity of each Newton sketch iteration scales as $\mathcal{O}(md^2)$, where m is at most d . In contrast, the standard unsketched form of the Newton update has complexity $\mathcal{O}(nd^2)$, so that the sketched method is computationally cheaper whenever there are more constraints than dimensions ($n > d$).

By increasing the barrier parameter τ , we obtain a sequence of solutions that approach the optimum to the LP, which we refer to as the central path. As a simple illustration, Figure 1 compares the central paths generated by the ordinary and sketched Newton updates for a polytope defined by $n = 32$ constraints in dimension $d = 2$. Each row shows three independent trials of the method for a given sketch dimension m ; the top, middle and bottom rows correspond to sketch dimensions $m \in \{d, 4d, 16d\}$ respectively. Note that as the sketch dimension m is increased, the central path taken by the sketched updates converges to the standard central path.

As a second example, we consider the problem of maximum likelihood estimation for generalized linear models.

Example 2 (Newton sketch for maximum likelihood estimation). The class of generalized linear models (GLMs) is used to model a wide variety of prediction and classification problems, in which the goal is to predict some output variable $y \in \mathcal{Y}$ on the basis of a covariate vector $a \in \mathbb{R}^d$. it includes as special cases the standard linear Gaussian model (in which $\mathcal{Y} = \mathbb{R}$), as well as logistic models for classification (in which $\mathcal{Y} = \{-1, +1\}$), as well as as **Poisson models** for count-valued responses (in which $\mathcal{Y} = \{0, 1, 2, \dots\}$). See the book [14] for further details and applications.

Given a collection of n observations $\{(y_i, a_i)\}_{i=1}^n$ of response-covariate pairs from some GLM, the problem of constrained maximum likelihood estimation be written in the form

$$\min_{x \in \mathcal{C}} \left\{ \underbrace{\sum_{i=1}^n \psi(\langle a_i, x \rangle, y_i)}_{f(x)} \right\}, \quad (10)$$

where $\psi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a given convex function, and $\mathcal{C} \subset \mathbb{R}^d$ is a convex constraint set, chosen by the user to enforce a certain type of structure in the solution. Important special cases of GLMs include the linear Gaussian model, in which $\psi(u, y) = \frac{1}{2}(y - u)^2$, and the problem (10)

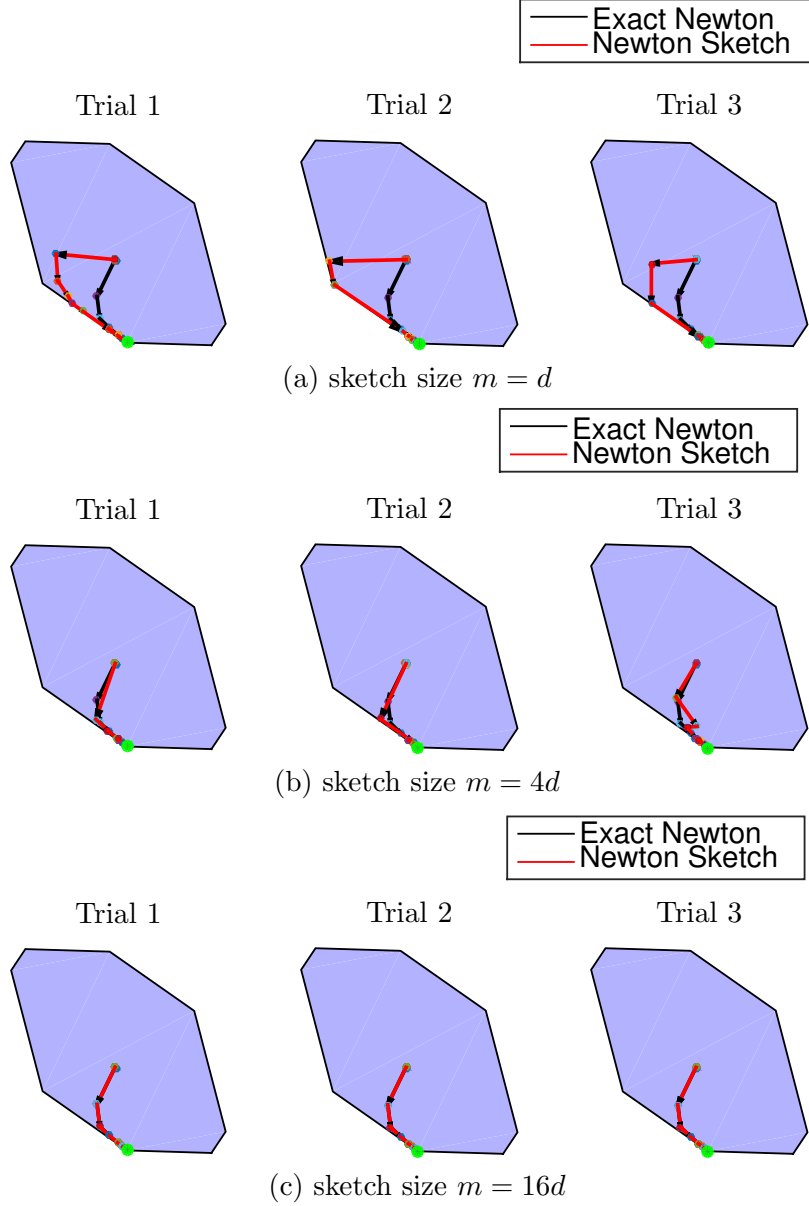


Figure 1. Comparisons of central paths for a simple linear program in two dimensions. Each row shows three independent trials for a given sketch dimension: across the rows, the sketch dimension ranges as $m \in \{d, 4d, 16d\}$. The black arrows show Newton steps taken by the standard interior point method, whereas red arrows show the steps taken by the sketched version. The green point at the vertex represents the optimum. In all cases, the sketched algorithm converges to the optimum, and as the sketch dimension m increases, the sketched central path converges to the standard central path.

corresponds to a regularized form of least-squares, as well as the problem of logistic regression, obtained by setting $\psi(u, y) = \log(1 + \exp(-yu))$.

Letting $A \in \mathbb{R}^{n \times d}$ denote the data matrix with $a_i \in \mathbb{R}^d$ as its i^{th} row, the Hessian of the objective (10) takes the form

$$\nabla^2 f(x) = A^T \text{diag}(\psi''(a_i^T x))_{i=1}^n A$$

Since the function ψ is convex, we are guaranteed that $\psi''(a_i^T x) \geq 0$, and hence the quantity $\text{diag}(\psi''(a_i^T x))^{1/2} A$ can be used as an $n \times d$ matrix square-root. We return to explore this class of examples in more depth in Section 5.1.

3.3 Local convergence analysis using strong convexity

Returning now to the general setting, we now begin by proving a local convergence guarantee for the sketched Newton updates. In particular, this theorem provides insight into how large the sketch dimension m must be in order to guarantee good local behavior of the sketched Newton algorithm.

This choice of sketch dimension is determined by geometry of the problem, in particular in terms of the tangent cone defined by the optimum. Given a constraint set \mathcal{C} and the minimizer $x^* := \arg \min_{x \in \mathcal{C}} f(x)$, the tangent cone at x^* is given by

$$\mathcal{K} := \{\Delta \in \mathbb{R}^d \mid x^* + t\Delta \in \mathcal{C} \text{ for some } t > 0\}. \quad (11)$$

Recalling the definition of the Gaussian width from Section 2.3, our first main result requires the sketch dimension to satisfy a lower bound of the form

$$m \geq \frac{c}{\epsilon^2} \max_{x \in \mathcal{C}} \mathcal{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K}), \quad (12)$$

where $\epsilon \in (0, 1)$ is a user-defined tolerance, and c is a universal constant. Since the Hessian square-root $\nabla^2 f(x)^{1/2}$ has dimensions $n \times d$, this squared Gaussian width is at most $\min\{n, d\}$. This worst-case bound is achieved for an unconstrained problem (in which case $\mathcal{K} = \mathbb{R}^d$), but the Gaussian width can be substantially smaller for constrained problems. See the example following Theorem 1 for an illustration.

In addition to this Gaussian width, our analysis depends on the cone-constrained eigenvalues of the Hessian $\nabla^2 f(x^*)$, which are defined as

$$\gamma = \inf_{z \in \mathcal{K} \cap \mathcal{S}^{d-1}} \langle z, \nabla^2 f(x^*) z \rangle, \quad \text{and} \quad \beta = \sup_{z \in \mathcal{K} \cap \mathcal{S}^{d-1}} \langle z, \nabla^2 f(x^*) z \rangle, \quad (13)$$

In the unconstrained case ($\mathcal{C} = \mathbb{R}^d$), we have $\mathcal{K} = \mathbb{R}^d$, and so that γ and β reduce to the minimum and maximum eigenvalues of the Hessian $\nabla^2 f(x^*)$. In the classical analysis of Newton's method, these quantities measure the strong convexity and smoothness parameters of the function f .

With this set-up, the following theorem is applicable to any twice-differentiable objective f with cone-constrained eigenvalues (γ, β) defined in equation (13), and with Hessian that is L -Lipschitz continuous, as defined in equation (2).

Theorem 1 (Local convergence of Newton Sketch). *For given parameters $\delta, \epsilon \in (0, 1)$, consider the Newton sketch updates (6) based on an initialization x^0 such that $\|x^0 - x^*\|_2 \leq \delta \frac{\gamma}{8L}$, and a sketch dimension m satisfying the lower bound (12). Then with probability at least $1 - c_1 e^{-c_2 m}$, the ℓ_2 -error satisfies the recursion*

$$\|x^{t+1} - x^*\|_2 \leq \epsilon \frac{\beta}{\gamma} \|x^t - x^*\|_2 + \frac{4L}{\gamma} \|x^t - x^*\|_2^2. \quad (14)$$

The bound (14) shows that when ϵ is set to a fixed constant—say $\epsilon = 1/4$ —the algorithm displays a linear-quadratic convergence rate in terms of the error $\Delta^t = x^t - x^*$. More

specifically, the rate is initially quadratic—that is, $\|\Delta^{t+1}\|_2 \approx \frac{4L}{\gamma} \|\Delta^t\|_2^2$ when $\|\Delta^t\|_2$ is large. However, as the iterations progress and $\|\Delta^t\|_2$ becomes substantially less than 1, then the rate becomes linear—meaning that $\|\Delta^{t+1}\|_2 \approx \epsilon \frac{\beta}{\gamma} \|\Delta^t\|_2$ —since the term $\frac{4L}{\gamma} \|\Delta^t\|_2^2$ becomes negligible compared to $\epsilon \frac{\beta}{\gamma} \|\Delta^t\|_2$. If we perform N steps in total, the linear rate guarantees the conservative error bounds

$$\|x^N - x^*\|_2 \leq \frac{\gamma}{8L} \left(\frac{1}{2} + \epsilon \frac{\beta}{\gamma} \right)^N, \quad \text{and} \quad f(x^N) - f(x^*) \leq \frac{\beta\gamma}{8L} \left(\frac{1}{2} + \epsilon \frac{\beta}{\gamma} \right)^N. \quad (15)$$

A notable feature of Theorem 1 is that, depending on the structure of the problem, the linear-quadratic convergence can be obtained using a sketch dimension m that is substantially smaller than $\min\{n, d\}$. As an illustrative example, we performed simulations for some instantiations of a portfolio optimization problem: it is a linearly-constrained quadratic program of the form

$$\min_{\substack{x \geq 0 \\ \sum_{j=1}^d x_j = 1}} \left\{ \frac{1}{2} x^T A^T A x - \langle c, x \rangle \right\}, \quad (16)$$

where $A \in \mathbb{R}^{n \times d}$ and $c \in \mathbb{R}^d$ are empirically estimated matrices and vectors (see Section 5.3 for more details). We used the Newton sketch to solve different sizes of this problem $d \in \{10, 20, 30, 40, 50, 60\}$, and with $n = d^3$ in each case. Each problem was constructed so that the optimum x^* had at most $s = \lceil 2 \log(d) \rceil$ non-zero entries. A calculation of the Gaussian width for this problem (see Appendix C for the details) shows that it suffices to take a sketch dimension $m \gtrsim s \log d$, and we implemented the algorithm with this choice. Figure 2

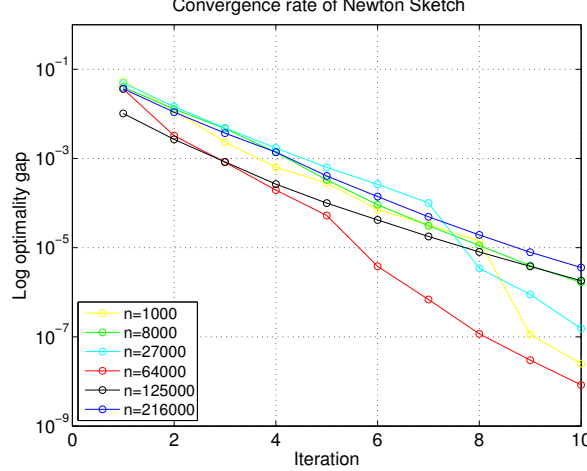


Figure 2. Empirical illustration of the linear convergence of the Newton sketch algorithm for an ensemble of portfolio optimization problems (16). In all cases, the algorithm was implemented using a sketch dimension $m = \lceil 4s \log d \rceil$, where s is an upper bound on the number of non-zeros in the optimal solution x^* ; this quantity satisfies the required lower bound (12), and consistent with the theory, the algorithm displays linear convergence.

shows the convergence rate of the Newton sketch algorithm for the six different problem sizes: consistent with our theory, the sketch dimension $m \ll \min\{d, n\}$ suffices to guarantee linear convergence in all cases.

It is also possible to obtain an asymptotically super-linear rate by using an iteration-dependent sketching accuracy $\epsilon = \epsilon(t)$. The following corollary summarizes one such possible guarantee:

Corollary 1. *Consider the Newton sketch iterates using the iteration-dependent sketching accuracy $\epsilon(t) = \frac{1}{\log(1+t)}$. Then with the same probability as in Theorem 1, we have*

$$\|x^{t+1} - x^*\|_2 \leq \frac{1}{\log(1+t)} \frac{\beta}{\gamma} \|x^t - x^*\|_2 + \frac{4L}{\gamma} \|x^t - x^*\|_2^2,$$

and consequently, super-linear convergence is obtained—namely, $\lim_{t \rightarrow \infty} \frac{\|x^{t+1} - x^*\|_2}{\|x^t - x^*\|_2} = 0$.

Note that the price for this super-linear convergence is that the sketch size is inflated by the factor $\epsilon^{-2}(t) = \log^2(1+t)$, so it is only logarithmic in the iteration number.

4 Newton sketch for self-concordant functions

The analysis and complexity estimates given in the previous section involve the curvature constants (γ, β) and the Lipschitz constant L , which are seldom known in practice. Moreover, as with the analysis of classical Newton method, the theory is local, in that the linear-quadratic convergence takes place once the iterates enter a suitable basin of the origin.

In this section, we seek to obtain global convergence results that do not depend on unknown problem parameters. As in the classical analysis, the appropriate setting in which to seek such results is for self-concordant functions, and using an appropriate form of backtracking line search. We begin by analyzing the unconstrained case, and then discuss extensions to constrained problems with self-concordant barriers. In each case, we show that given a suitable lower bound on the sketch dimension, the sketched Newton updates can be equipped with global convergence guarantees that hold with exponentially high probability. Moreover, the total number of iterations does not depend on any unknown constants such as strong convexity and Lipschitz parameters.

4.1 Unconstrained case

In this section, we consider the unconstrained optimization problem $\min_{x \in \mathbb{R}^d} f(x)$, where f is a closed convex self-concordant function which is bounded below. Note that a closed convex function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is *self-concordant* if

$$|\phi'''(x)| \leq 2 (\phi''(x))^{3/2}. \quad (17)$$

This definition can be extended to a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by imposing this requirement on the univariate functions $\phi_{x,y}(t) := f(x + ty)$, for all choices of x, y in the domain of f . Examples of self-concordant functions include linear and quadratic functions and negative logarithm. Self concordance is preserved under addition and affine transformations.

Our main result provide a bound on the total number of Newton sketch iterations required to obtain a δ -accurate solution without imposing any sort of initialization condition (as was done in our previous analysis). This bound scales proportionally to $\log(1/\delta)$ and inversely in a parameter ν that depends on sketching accuracy $\epsilon \in (0, \frac{1}{4})$ and backtracking parameters (a, b) via

$$\nu = ab \frac{\eta^2}{1 + (\frac{1+\epsilon}{1-\epsilon})\eta} \quad \text{where} \quad \eta = \frac{1}{8} \frac{1 - \frac{1}{2}(\frac{1+\epsilon}{1-\epsilon})^2 - a}{(\frac{1+\epsilon}{1-\epsilon})^3}. \quad (18)$$

Algorithm 1 Unconstrained Newton Sketch with backtracking line search

Input: Starting point x^0 , tolerance $\delta > 0$, (a, b) line-search parameters, sketching matrices $\{S^t\}_{t=0}^\infty \in \mathbb{R}^{m \times n}$.
 1: Compute approximate Newton step Δx^t and approximate Newton decrement $\lambda(x)$

$$\Delta x^t := \arg \min_{\Delta} \langle \nabla f(x^t), \Delta \rangle + \frac{1}{2} \|S^t(\nabla^2 f(x^t))^{1/2} \Delta\|_2^2;$$

$$\tilde{\lambda}_f(x^t) := \nabla f(x^t)^T \Delta x^t.$$

2: Quit if $\tilde{\lambda}(x^t)^2/2 \leq \delta$.

3: Line search: choose μ : **while** $f(x^t + \mu \Delta x^t) > f(x^t) + a\mu \lambda(x^t)$, $\mu \leftarrow b\mu$

4: Update: $x^{t+1} = x^t + \mu \Delta x^t$

Output: minimizer x^t , optimality gap $\lambda(x^t)$

Theorem 2. *Let f be a strictly convex self-concordant function. Given a sketching matrix $S \in \mathbb{R}^{m \times n}$ with $m \geq \frac{c_3}{c_2} \max_{x \in \mathcal{C}} \text{rank}(\nabla^2 f(x)) = \frac{c_3}{c_2} d$, the number of total iterations T for obtaining an δ approximate solution in function value via Algorithm 1 is bounded by*

$$T = \frac{f(x^0) - f(x^*)}{\nu} + 0.65 \log_2 \left(\frac{1}{16\delta} \right),$$

with probability at least $1 - c_1 N e^{-c_2 m}$.

The bound in the above theorem shows that the convergence of the Newton Sketch is independent of the properties of the function f and problem parameters, similar to classical Newton's method. Note that for problems with $n > d$, the complexity of each Newton sketch step is at most $\mathcal{O}(d^3 + nd \log d)$, which is smaller than that of Newton's Method ($\mathcal{O}(nd^2)$), and also smaller than typical first-order optimization methods ($\mathcal{O}(nd)$) whenever $n > d^2$.

4.2 Newton Sketch with self-concordant barriers

We now turn to the more general constrained case. Given a closed, convex self-concordant function $f_0 : \mathbb{R}^d \rightarrow \mathbb{R}$, let \mathcal{C} be a convex subset of \mathbb{R}^d , and consider the constrained optimization problem $\min_{x \in \mathcal{C}} f_0(x)$. If we are given a convex self-concordant barrier function g for the constraint set \mathcal{C} , it is equivalent to consider the unconstrained problem

$$\min_{x \in \mathbb{R}^d} \left\{ \underbrace{f_0(x) + g(x)}_{f(x)} \right\}.$$

One way in which to solve this unconstrained problem is by sketching the Hessian of both f_0 and g , in which case the theory of the previous section is applicable. However, there are many cases in which the constraints describing \mathcal{C} are relatively simple, and so the Hessian of g is highly-structured. For instance, if the constraint set is the usual simplex (i.e., $x \geq 0$ and $\langle 1, x \rangle \leq 1$), then the Hessian of the associated log barrier function is a diagonal matrix plus a rank one matrix. Other examples include problems for which g has a separable structure; such functions frequently arise as regularizers for ill-posed inverse problems. Examples of such regularizers include ℓ_2 regularization $g(x) = \frac{1}{2} \|x\|_2^2$, graph regularization $g(x) = \frac{1}{2} \sum_{i,j \in E} (x_i - x_j)^2$ induced by an edge set E (e.g., finite differences) and also other differentiable norms $g(x) = \left(\sum_{i=1}^d x_i^p \right)^{1/p}$ for $1 < p < \infty$.

In all such cases, an attractive strategy is to apply a *partial Newton sketch*, in which we sketch the Hessian term $\nabla^2 f_0(x)$ and retain the exact Hessian $\nabla^2 g(x)$, as in the previously

described updates (8). More formally, Algorithm 2 provides a summary of the steps, including the choice of the line search parameters. The main result of this section provides a guarantee on this algorithm, assuming that the sequence of sketch dimensions $\{m^t\}_{t=0}^\infty$ is appropriately chosen.

Algorithm 2 Newton Sketch with self-concordant barriers

Input: Starting point x^0 , constraint \mathcal{C} , corresponding barrier function g such that $f = f_0 + g$, tolerance $\delta > 0$, (α, β) line-search parameters, sketching matrices $S^t \in \mathbb{R}^{m \times n}$.

1: Compute approximate Newton step Δx^t and approximate Newton decrement $\lambda(x)$.

$$\begin{aligned}\Delta x^t &:= \arg \min_{x^t + \Delta \in \mathcal{C}} \langle \nabla f(x^t), \Delta \rangle + \frac{1}{2} \|S^t (\nabla^2 f_0(x^t))^{1/2} \Delta\|_2^2 + \frac{1}{2} \Delta^T \nabla^2 g(x^t) \Delta; \\ \tilde{\lambda}_f(x^t) &:= \nabla f(x^t)^T \Delta x^t\end{aligned}$$

2: Quit if $\tilde{\lambda}(x^t)^2/2 \leq \delta$.

3: Line search: choose μ : **while** $f(x^t + \mu \Delta x^t) > f(x^t) + \alpha \mu \lambda(x^t)$, $\mu \leftarrow \beta \mu$.

4: Update: $x^{t+1} = x^t + \mu \Delta x^t$.

Output: minimizer x^t , optimality gap $\lambda(x^t)$.

The choice of sketch dimensions depends on the tangent cones defined by the iterates, namely the sets

$$\mathcal{K}^t := \{\Delta \in \mathbb{R}^d \mid x^t + \alpha \Delta \in \mathcal{C} \text{ for some } \alpha > 0\}.$$

For a given sketch accuracy $\epsilon \in (0, 1)$, we require that the sequence of sketch dimensions satisfies the lower bound

$$m^t \geq \frac{c_3}{\epsilon^2} \max_{x \in \mathcal{C}} \mathcal{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K}^t). \quad (19)$$

Finally, the reader should recall the parameter ν was defined in equation (18), which depends only on the sketching accuracy ϵ and the line search parameters. Given this set-up, we have the following guarantee:

Theorem 3. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex and self-concordant function, and let $g : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex and self-concordant barrier for the convex set \mathcal{C} . Suppose that we implement Algorithm 2 with sketch dimensions $\{m^t\}_{t \geq 0}$ satisfying the lower bound (19). Then taking*

$$N = \frac{f(x^0) - f(x^*)}{\nu} + 0.65 \log_2\left(\frac{1}{16\delta}\right) \quad \text{iterations,}$$

suffices to obtain δ -approximate solution in function value with probability at least $1 - c_1 N e^{-c_2 m}$.

Thus, we see that the Newton Sketch method can also be used with self-concordant barrier functions, which considerably extends its scope. Section 5.5 provides a numerical illustration of its performance in this context. As we discuss in the next section, there is a flexibility in choosing the decomposition f_0 and g corresponding to objective and barrier, which enables us to also sketch the constraints.

4.3 Sketching with interior point methods

In this section, we discuss the application of Newton Sketch to a form of barrier or interior point methods. In particular we discuss two different strategies and provide rigorous worst-case complexity results when the functions in the objective and constraints are self-concordant.

Algorithm 3 Interior point methods using Newton Sketch

Input: Strictly feasible starting point x^0 , initial parameter τ^0 s.t. $\tau := \tau^0 > 0$, $\mu > 1$, tolerance $\delta > 0$.

- 1: Centering step: Compute $\hat{x}(\tau)$ by Newton Sketch with backtracking line-search initialized at x using Algorithm 1 or Algorithm 2.
- 2: Update $x := \hat{x}(\tau)$.
- 3: Quit if $r/\tau \leq \delta$.
- 4: Increase τ by $\tau := \mu\tau$.

Output: minimizer $\hat{x}(\tau)$.

More precisely, let us consider a problem of the form

$$\min_{x \in \mathbb{R}^d} f_0(x) \quad \text{subject to} \quad g_j(x) \leq 0 \quad \text{for } j = 1, \dots, r, \quad (20)$$

where f_0 and $\{g_j\}_{j=1}^r$ are twice-differentiable convex functions. We assume that there exists a unique solution x^* to the above problem.

The barrier method for computing x^* is based on solving a sequence of problems of the form

$$\hat{x}(\tau) := \arg \min_{x \in \mathbb{R}^d} \left\{ \tau f_0(x) - \sum_{j=1}^r \log(-g_j(x)) \right\}, \quad (21)$$

for increasing values of the parameter $\tau \geq 1$. The family of solutions $\{\hat{x}(\tau)\}_{\tau \geq 1}$ trace out what is known as the central path. A standard bound (e.g., [4]) on the sub-optimality of $\hat{x}(\tau)$ is given by

$$f_0(\hat{x}(\tau)) - f_0(x^*) \leq \frac{r}{\tau}.$$

The barrier method successively updates the penalty parameter τ and also the starting points supplied to Newton's method using previous solutions.

Since Newton's method lies at the heart of the barrier method, we can obtain a fast version by replacing the exact Newton minimization with the Newton sketch. Algorithm 3 provides a precise description of this strategy. As noted in Step 1, there are two different strategies in dealing with the convex constraints $g_j(x) \leq 0$ for $j = 1, \dots, r$:

- *Full sketch:* Sketch the full Hessian of the objective function (21) using Algorithm 1,
- *Partial sketch:* Sketch only the Hessians corresponding to a subset of the functions $\{f_0, g_j, j = 1, \dots, r\}$, and use exact Hessians for the other functions. Apply Algorithm 2.

As shown by our theory, either approach leads to the same convergence guarantees, but the associated computational complexity can vary depending both on how data enters the objective and constraints, as well as the Hessian structure arising from particular functions. The following theorem is an application of the classical results on the barrier method tailored for Newton Sketch using any of the above strategies (see e.g., [4]). As before, the key parameter ν was defined in Theorem 2.

Theorem 4 (Newton Sketch complexity for interior point methods). *For a given target accuracy $\delta \in (0, 1)$ and any $\mu > 1$, the total number of Newton Sketch iterations required to obtain a δ -accurate solution using Algorithm 3 is at most*

$$\left\lceil \frac{\log(r/(\tau^0 \delta))}{\log \mu} \right\rceil \left(\frac{r(\mu - 1 - \log \mu)}{\gamma} + 0.65 \log_2\left(\frac{1}{16\delta}\right) \right). \quad (22)$$

If the parameter μ is set to minimize the above upper-bound, the choice $\mu = 1 + \frac{1}{r}$ yields $\mathcal{O}(\sqrt{r})$ iterations. However, when applying the standard Newton method, this “optimal” choice is typically not used in practice: instead, it is common to use a fixed value of $\mu \in [2, 100]$. In experiments, experience suggests that the number of Newton iterations needed is a constant independent of r and other parameters. Theorem 4 allows us to obtain faster interior point solvers with rigorous worst-case complexity results. We show different applications of Algorithm 3 in the following section.

5 Applications and numerical results

In this section, we discuss some applications of the Newton sketch to different optimization problems. In particular, we show various forms of Hessian structure that arise in applications, and how the Newton sketch can be computed. When the objective and/or the constraints contain more than one term, the barrier method with Newton Sketch has some flexibility in sketching. We discuss the choices of partial Hessian sketching strategy in the barrier method. It is also possible to apply the sketch in the primal or dual form, and we provide illustrations of both strategies here.

5.1 Estimation in generalized linear models

Recall the problem of (constrained) maximum likelihood estimation for a generalized linear model, as previously introduced in Example 2. It leads to the family of optimization problems (10): here $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is a given convex function arising from the probabilistic model, and $\mathcal{C} \subseteq \mathbb{R}^d$ is a closed convex set that is used to enforce a certain type of structure in the solution. Popular choices of such constraints include ℓ_1 -balls (for enforcing sparsity in a vector), nuclear norms (for enforcing low-rank structure in a matrix), and other non-differentiable semi-norms based on total variation (e.g., $\sum_{j=1}^{d-1} |x_{j+1} - x_j|$), useful for enforcing smoothness or clustering constraints.

Suppose that we apply the Newton sketch algorithm to the optimization problem (10). Given the current iterate x^t , computing the next iterate x^{t+1} requires solving the constrained quadratic program

$$\min_{x \in \mathcal{C}} \left\{ \frac{1}{2} \|S \text{diag}(\psi''(\langle a_i, x^t \rangle, y_i))^{1/2} A(x - x^t)\|_2^2 + \sum_{i=1}^n \langle x, \psi'(\langle a_i, x^t \rangle, y_i) \rangle \right\}. \quad (23)$$

When the constraint \mathcal{C} is a scaled version of the ℓ_1 -ball—that is, $\mathcal{C} = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq R\}$ for some radius $R > 0$ —the convex program (23) is an instance of the Lasso program [23], for which there is a very large body of work. For small values of R , where the cardinality of the solution x is very small, an effective strategy is to apply a homotopy type algorithm, also known as LARS [7, 9], which solves the optimality conditions starting from $R = 0$. For other sets \mathcal{C} , another popular choice is projected gradient descent, which is efficient when projection onto \mathcal{C} is computationally simple.

Focusing on the ℓ_1 -constrained case, let us consider the problem of choosing a suitable sketch dimension m . Our choice involves the ℓ_1 -restricted minimal eigenvalue of the data matrix $A^T A$, which is given by

$$\gamma_s^-(A) := \min_{\substack{\|z\|_2=1 \\ \|z\|_1 \leq 2\sqrt{s}}} \|Az\|_2^2. \quad (24)$$

Note that we are always guaranteed that $\gamma_s^-(A) \geq \lambda_{\min}(A^T A)$. It also involves certain quantities that depend on the function ψ , namely

$$\psi''_{\min} := \min_{x \in \mathcal{C}} \min_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i), \quad \text{and} \quad \psi''_{\max} := \max_{x \in \mathcal{C}} \max_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i),$$

where $a_i \in \mathbb{R}^d$ is the i^{th} row of A . With this set-up, supposing that the optimal solution x^* has cardinality at most $\|x^*\|_0 \leq s$, then it can be shown (see Lemma 8 in Appendix C) that it suffices to take a sketch size

$$m = c_0 \frac{\psi''_{\max}}{\psi''_{\min}} \frac{\max_{j=1, \dots, d} \|A_j\|_2^2}{\gamma_s^-(A)} s \log d, \quad (25)$$

where c_0 is a universal constant. Let us consider some examples to illustrate:

- Least-Squares regression: $\psi(u) = \frac{1}{2}u^2$, $\psi''(u) = 1$ and $\psi''_{\min} = \psi''_{\max} = 1$.
- Poisson regression: $\psi(u) = e^u$, $\psi''(u) = e^u$ and $\frac{\psi''_{\max}}{\psi''_{\min}} = \frac{e^{RA_{\max}}}{e^{-RA_{\min}}}$
- Logistic regression: $\psi(u) = \log(1+e^u)$, $\psi''(u) = \frac{e^u}{(e^u+1)^2}$ and $\frac{\psi''_{\max}}{\psi''_{\min}} = \frac{e^{RA_{\min}}}{e^{-RA_{\max}}} \frac{(e^{-RA_{\max}}+1)^2}{(e^{RA_{\min}}+1)^2}$,

where $A_{\max} := \max_{i=1, \dots, n} \|a_i\|_{\infty}$, and $A_{\min} := \min_{i=1, \dots, n} \|a_i\|_{\infty}$.

For typical distributions of the data matrices, the sketch size choice given in equation (25) is $\mathcal{O}(s \log d)$. As an example, consider data matrices $A \in \mathbb{R}^{n \times d}$ where each row is independently sampled from a sub-Gaussian distribution with variance 1. Then standard results on random matrices [24] show that $\gamma_s^-(A) > 1/2$ as long as $n > c_1 s \log d$ for a sufficiently large constant c_1 . In addition, we have $\max_{j=1, \dots, d} \|A_j\|_2^2 = \mathcal{O}(n)$, as well as $\frac{\psi''_{\max}}{\psi''_{\min}} = \mathcal{O}(\log(n))$. For such problems, the per iteration complexity of Newton Sketch update scales as $\mathcal{O}(s^2 d \log^2(d))$ using standard Lasso solvers (e.g., [11]) or as $\mathcal{O}(sd \log(d))$ using projected gradient descent. Both of these scalings are substantially smaller than conventional algorithms that fail to exploit the small intrinsic dimension of the tangent cone.

5.2 Semidefinite programs

The Newton sketch can also be applied to semidefinite programs. As one illustration, let us consider the metric learning problem studied in machine learning. Given feature vectors $a_1, \dots, a_n \in \mathbb{R}^d$ and corresponding indicator $y_{ij} \in \{-1, +1\}^n$ where $y_{ij} = +1$ if a_i and a_j belong to the same label and $y_{ij} = -1$ otherwise for all $i \neq j$ and $1 \leq i, j \leq n$. The task is to learn a positive semidefinite matrix X which represents a metric such that the semi-norm $\|a\|_X := \sqrt{\langle a, Xa \rangle}$ establishes a nearness measure depending on class label. Using ℓ_2 -loss, the optimization can be stated as the following semi-definite program (SDP)

$$\min_{X \succeq 0} \left\{ \sum_{i \neq j}^{\binom{n}{2}} (\langle X, (a_i - a_j)(a_i - a_j)^T \rangle - y_{ij})^2 + \lambda \text{trace}(X) \right\}.$$

Here the term $\text{trace}(X)$, along with its multiplicative pre-factor $\lambda > 0$ that can be adjusted by the user, is a regularization term for encouraging a relatively low-rank solution. Using

the standard self-concordant barrier $X \mapsto \log \det(X)$ for the PSD cone, the barrier method involves solving a sequence of sub-problems of the form

$$\min_{X \in \mathbb{R}^{d \times d}} \left\{ \underbrace{\tau \sum_{i=1}^n (\langle X, a_i a_i^T \rangle - y_i)^2 + \tau \lambda \text{trace } X - \log \det(X)}_{f(\text{vec}(X))} \right\}.$$

Now the Hessian of the function $\text{vec}(X) \mapsto f(\text{vec}(X))$ is a $d^2 \times d^2$ matrix given by

$$\nabla^2 f(\text{vec}(X)) = \tau \sum_{i \neq j}^{\binom{n}{2}} \text{vec}(A_{ij}) \text{vec}(A_{ij})^T + X^{-1} \otimes X^{-1},$$

where $A_{ij} := (a_i - a_j)(a_i - a_j)^T$. Then we can apply the barrier method with partial Hessian sketch on the first term, $\{S_{ij} \text{vec}(A_{ij})\}_{i \neq j}$ and exact Hessian for the second term. Since the vectorized decision variable is $\text{vec}(X) \in \mathbb{R}^{d^2}$ the complexity of Newton Sketch is $\mathcal{O}(m^2 d^2)$ while the complexity of a classical SDP interior-point solver is $\mathcal{O}(nd^4)$.

5.3 Portfolio optimization and SVMs

Here we consider the Markowitz formulation of the portfolio optimization problem [13]. The objective is to find $x \in \mathbb{R}^d$ belonging to the unit simplex, which corresponds to non-negative weights associated with each of d possible assets, so as to maximize the expected return minus a coefficient times the variance of the return. Letting $\mu \in \mathbb{R}^d$ denote a vector corresponding to mean return of the assets, and we let $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric, positive semidefinite matrix, covariance of the returns. The optimization problem is given by

$$\max_{x \geq 0, \sum_{j=1}^d x_j \leq 1} \left\{ \langle \mu, x \rangle - \lambda x^T \Sigma x \right\}. \quad (26)$$

The covariance of returns is often estimated from past stock data via empirical covariance, $\Sigma = A^T A$ where the columns of A are time series corresponding to assets normalized by \sqrt{n} , where n is the length of the observation window.

The barrier method can be used solve the above problem by solving penalized problems of the

$$\min_{x \in \mathbb{R}^d} \left\{ \underbrace{-\tau \mu^T x + \tau \lambda x^T A^T A x - \sum_{i=1}^d \log(\langle e_i, x \rangle) - \log(1 - \langle 1, x \rangle)}_{f(x)} \right\},$$

where $e_i \in \mathbb{R}^d$ is the i^{th} element of the canonical basis and 1 is row vector of all-ones. Then the Hessian of the above barrier penalized formulation can be written as

$$\nabla^2 f(x) = \tau \lambda A^T A + \text{diag}(x_i^2)^{-1} + 11^T$$

Consequently we can sketch the data dependent part of the Hessian via $\tau \lambda S A$ which has at most rank m and keep the remaining terms in the Hessian exact. Since the matrix 11^T is rank one, the resulting sketched estimate is therefore diagonal plus rank $(m+1)$ where the matrix inversion lemma can be applied for efficient computation of the Newton Sketch update (see e.g. [8]). Therefore, as long as $m \leq d$, the complexity per iteration scales as $\mathcal{O}(md^2)$, which is cheaper than the $\mathcal{O}(nd^2)$ per step complexity associated with classical interior point methods. We also note that support vector machine classification problems with squared hinge loss also has the same form as in (26) (see e.g. [20]) where the same strategy can be applied.

5.4 Unconstrained logistic regression with $d \ll n$

Let us now turn to some numerical comparisons of the Newton Sketch with other popular optimization methods for large-scale instances of logistic regression. More specifically, we generated a feature matrix $A \in \mathbb{R}^{n \times d}$ based on $d = 100$ features and $n = 16384$ observations. Each row $a_i \in \mathbb{R}^d$ was generated from the d -variate Gaussian distribution $N(0, \Sigma)$ where $\Sigma_{ij} = 2|0.99|^{i-j}$. As shown in Figure 3, the convergence of the algorithm per iteration is very similar to Newton's method. Besides the original Newton's method, the other algorithms compared are

- Gradient Descent (GD) with backtracking line search
- Accelerated Gradient Descent (Acc. GD) adapted for strongly convex functions with manually tuned parameters.
- Stochastic Gradient Descent (SGD) with the classical step size choice $1/\sqrt{t}$
- Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) approximating the Hessian with gradients.

For each problem, we averaged the performance of the randomized algorithms (Newton sketch and SGD) over 10 independent trials. We ran the Newton sketch algorithm with sketch size $m = 6d$. To be fair in comparisons, we performed hand-tuning of the stepsize parameters in the gradient-based methods so as to optimize their performance. The top panel in Figure 3 plots the log duality gap versus the number of iterations: as expected, on this scale, the classical form of Newton's method is the fastest, whereas the SGD method is the slowest. However, when the log optimality gap is plotted versus the wall-clock time in the bottom panel, we now see that the Newton sketch is the fastest.

5.5 A dual example: Lasso with $d \gg n$

The regularized Lasso problem takes the form $\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1 \right\}$, where $\lambda > 0$ is a user-specified regularization parameter. In this section, we consider efficient sketching strategies for this class of problems in the regime $d \gg n$. In particular, let us consider the corresponding dual program, given by

$$\max_{\|A^T w\|_\infty \leq \lambda} \left\{ -\frac{1}{2} \|y - w\|_2^2 \right\}.$$

By construction, the number of constraints d in the dual program is larger than the number of optimization variables n . If we apply the barrier method to solve this dual formulation, then we need to solve a sequence of problems of the form

$$\min_{w \in \mathbb{R}^n} \left\{ \underbrace{\tau \|y - w\|_2^2 - \sum_{j=1}^d \log(\lambda - \langle A_j, w \rangle) - \sum_{j=1}^d \log(\lambda + \langle A_j, w \rangle)}_{f(w)} \right\},$$

where $A_j \in \mathbb{R}^n$ denotes the j^{th} column of A . The Hessian of the above barrier penalized formulation can be written as

$$\nabla^2 f(w) = \tau I_n + \text{Adiag} \left(\frac{1}{(\lambda - \langle A_j, w \rangle)^2} \right) A^T + \text{Adiag} \left(\frac{1}{(\lambda + \langle A_j, w \rangle)^2} \right) A^T,$$

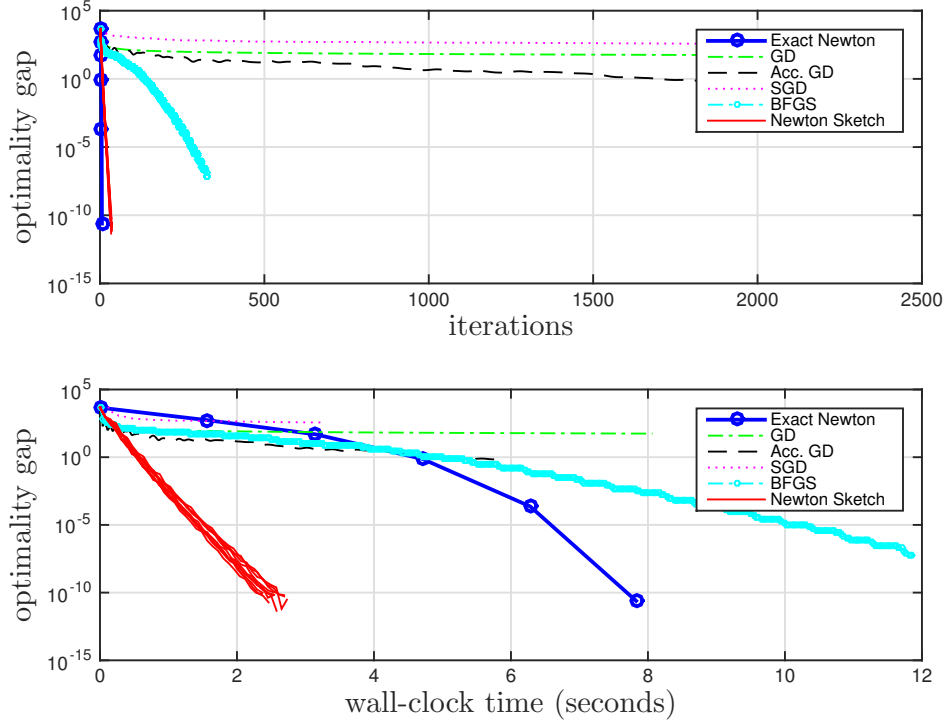


Figure 3. Newton Sketch algorithm outperforms other popular optimization methods. Plots of the log optimality gap versus iteration number (top) and plots of the log optimality gap versus wall-clock time (bottom). Newton Sketch empirically provides the best accuracy in smallest wall-clock time, and does not require knowledge of problem-dependent quantities (such as strong convexity and smoothness parameters).

Consequently we can keep the first term in the Hessian, τI exact and apply partial sketching to the Hessians of the last two terms via

$$S \text{diag} \left(\frac{1}{|\lambda - \langle A_j, w \rangle|} + \frac{1}{|\lambda + \langle A_j, w \rangle|} \right) A^T.$$

Since the partially sketched Hessian is of the form $tI_n + VV^T$, where V is rank at most m , we can use matrix inversion lemma for efficiently calculating Newton Sketch updates. The complexity of the above strategy for $d > n$ is $\mathcal{O}(dm^2)$, where m is at most d , whereas traditional interior point solvers are typically $\mathcal{O}(dn^2)$ per iteration.

In order to test this algorithm, we generated a feature matrix $A \in \mathbb{R}^{n \times d}$ with $d = 4096$ features and $n = 50$ observations. Each row $a_i \in \mathbb{R}^d$ was generated from the multivariate Gaussian distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 2 * |0.99|^{i-j}$. For a given problem instance, we ran 10 independent trials of the sketched barrier method, and compared the results to the original barrier method. Figure 4 plots the the duality gap versus iteration number (top panel) and versus the wall-clock time (bottom panel) for the original barrier method (blue) and sketched barrier method (red): although the sketched algorithm requires more iterations, these iterations are cheaper, leading to a smaller wall-clock time. This point is reinforced by Figure 5, where we plot the wall-clock time required to reach a duality gap of 10^{-6} versus the number of features n in problem families of increasing size. Note that the sketched barrier

method outperforms the original barrier method, with significantly less computation time for obtaining similar accuracy.

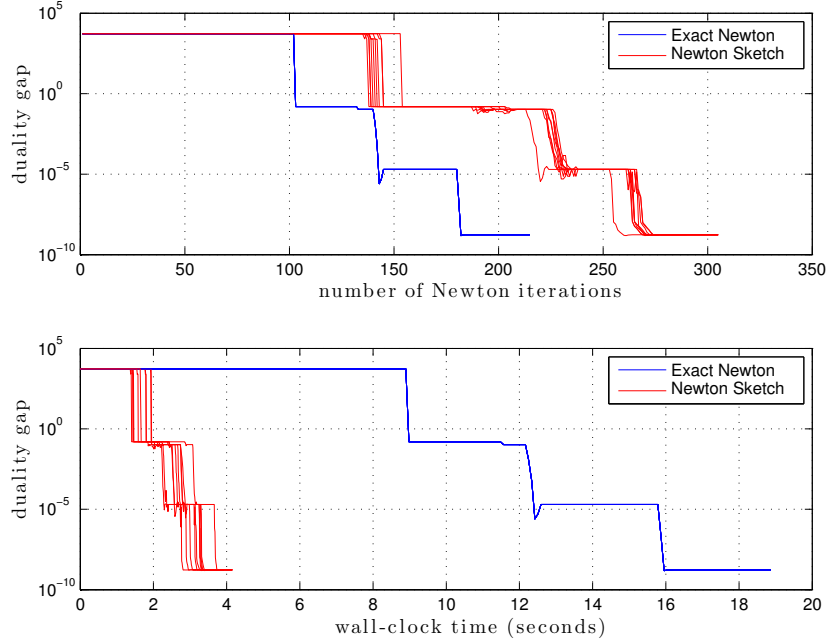


Figure 4. Plots of the duality gap versus iteration number (top panel) and duality gap versus wall-clock time (bottom panel) for the original barrier method (blue) and sketched barrier method (red). The sketched interior point method is run 10 times independently yielding slightly different curves in red. While the sketched method requires more iterations, its overall wall-clock time is much smaller.

6 Proofs

We now turn to the proofs of our theorems, with more technical details deferred to the appendices.

6.1 Proof of Theorem 1

Throughout this proof, we let $r \in \mathcal{S}^{d-1}$ denote a fixed vector that is independent of the sketch matrix S^t and the current iterate x^t . We then define the following pair of random variables

$$\begin{aligned} Z_1(S; x) &:= \sup_{w \in \nabla^2 f(x)^{1/2} \mathcal{K} \cap \mathcal{S}^{n-1}} \langle w, (S^T S - I)r \rangle, \\ Z_2(S; x) &:= \inf_{w \in \nabla^2 f(x)^{1/2} \mathcal{K} \cap \mathcal{S}^{n-1}} \|Sw\|_2^2. \end{aligned}$$

These random variables are significant, because the core of our proof is based on establishing that the error vector $\Delta^t = x^t - x^*$ satisfies the recursive bound

$$\|\Delta^{t+1}\|_2 \leq \frac{6\beta Z_1^t}{\gamma Z_2^t} \|\Delta^t\|_2 + \frac{4L}{\gamma Z_2^t} \|\Delta^t\|_2^2, \quad (27)$$

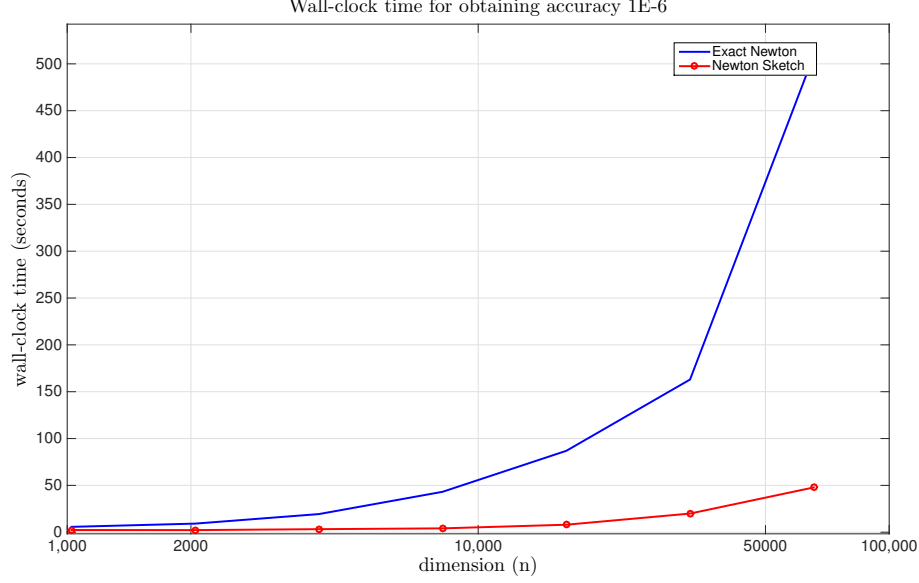


Figure 5. Plot of the wall-clock time in seconds for reaching a duality gap of 10^{-6} for the standard and sketched interior point methods as n increases (in log-scale). The sketched interior point method has significantly lower computation time compared to the original method.

where $Z_1^t := Z_1(S^t; x^t)$ and $Z_2^t := Z_2(S^t; x^t)$. We then combine this recursion with the following probabilistic guarantee on Z_1^t and Z_2^t . For a given tolerance parameter $\epsilon \in (0, \frac{1}{2}]$, consider the "good event"

$$\mathcal{E}^t := \left\{ Z_1^t \leq \frac{\epsilon}{2}, \text{ and } Z_2^t \geq 1 - \epsilon \right\}. \quad (28)$$

Lemma 1 (Sufficient conditions on sketch dimension [20]).

- (a) For sub-Gaussian sketch matrices, given a sketch size $m > \frac{c_0}{\epsilon^2} \max_{x \in \mathcal{C}} \mathcal{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K})$, we have

$$\mathbb{P}[\mathcal{E}^t] \geq 1 - c_1 e^{-c_2 m \epsilon^2}. \quad (29)$$

- (b) For randomized orthogonal system (ROS) sketches over the class of self-bounding cones, given a sketch size $m > \frac{c_0 \log^4 n}{\epsilon^2} \max_{x \in \mathcal{C}} \mathcal{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K})$, we have

$$\mathbb{P}[\mathcal{E}^t] \geq 1 - c_1 e^{-c_2 \frac{m \epsilon^2}{\log^4 n}}. \quad (30)$$

Combining Lemma 1 with the recursion (27) and re-scaling ϵ appropriately yields the claim of the theorem.

Accordingly, it remains to prove the recursion (27), and we do so via a basic inequality argument. Recall the function $x \mapsto \Phi(x; S^t)$ that underlies the sketch Newton update (6): since x^t and x^* are optimal and feasible for the constrained optimization problem, we have

$\Phi(x; S^t) \leq \Phi(x^*; S^t)$. Introducing the error vector $\Delta^t := x^t - x^*$, some straightforward algebra then leads to the *basic inequality*

$$\frac{1}{2} \|S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2 \leq \langle S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}, S \nabla^2 f(x^t)^{1/2} \Delta^t \rangle - \langle \nabla f(x^t) - \nabla f(x^*), \Delta^{t+1} \rangle \quad (31)$$

Let us first upper bound the right-hand side. By using the integral form of Taylor's expansion, we have $\langle \nabla f(x^t) - \nabla f(x^*), \Delta^{t+1} \rangle = \int_0^1 \langle \nabla^2 f(x^t + u(x^* - x^t)) \Delta^t, \Delta^{t+1} \rangle du$, and hence

$$\text{RHS} = \int_0^1 \langle [\nabla^2 f(x^t)^{1/2} (S^t)^T S^t \nabla^2 f(x^t)^{1/2} - \nabla^2 f(x^t + u(x^* - x^t))] \Delta^t, \Delta^{t+1} \rangle du$$

By adding and subtracting terms and then applying triangle inequality, we have the bound $\text{RHS} \leq T_1 + T_2$, where

$$\begin{aligned} T_1 &:= \left| \int_0^1 \langle [\nabla^2 f(x^t)^{1/2} ((S^t)^T S^t - I) \nabla^2 f(x^t)^{1/2}] \Delta^t, \Delta^{t+1} \rangle \right|, \quad \text{and} \\ T_2 &:= \int_0^1 \|\nabla^2 f(x^t + u(x^* - x^t)) - \nabla^2 f(x^t)\|_{\text{op}} du \|\Delta^t\|_2 \|\Delta^{t+1}\|_2. \end{aligned}$$

Now observe that the vector $r := \nabla^2 f(x^t)^{1/2} \Delta^t$ is independent of the randomness in S^t , whereas the vector $\nabla^2 f(x^t)^{1/2} \Delta^{t+1}$ belongs to the cone $\nabla^2 f(x^t)^{1/2} \mathcal{K}$. Consequently, by the definition of Z_1 , we have

$$T_1 \leq Z_1^t \|\nabla^2 f(x^t)^{1/2} \Delta^t\|_2 \|\nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2. \quad (32)$$

Now note that using the fact that β controls the smoothness of the gradient and the Lipschitz continuity of Hessian we can upper bound the terms on the above right-hand side as follows

$$\begin{aligned} \langle \Delta^t, \nabla^2 f(x^*) \Delta^t \rangle &= \langle \Delta^t, \nabla^2 f(x^*) \Delta^t \rangle + \langle \Delta^t, (\nabla^2 f(x^t) - \nabla^2 f(x^*)) \Delta^t \rangle \\ &\leq \{\beta + L \|\Delta^t\|_2\} \|\Delta^t\|_2^2, \end{aligned}$$

and similarly, $\langle \Delta^{t+1}, \nabla^2 f(x^*) \Delta^{t+1} \rangle \leq \{\beta + L \|\Delta^t\|_2\} \|\Delta^{t+1}\|_2^2$. Combining the above bounds with (32) we obtain

$$T_1 \leq Z_1^t \{\beta + L \|\Delta^t\|_2\} \|\Delta^{t+1}\|_2 \|\Delta^t\|_2. \quad (33)$$

On the other hand, by the L -Lipschitz condition on the Hessian, we have

$$T_2 \leq L \|\Delta^t\|_2^2 \|\Delta^{t+1}\|_2.$$

Substituting these two bounds into our basic inequality, we have

$$\frac{1}{2} \|S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2 \leq Z_1^t \{\beta + L \|\Delta^t\|_2\} \|\Delta^t\|_2 \|\Delta^{t+1}\|_2 + L \|\Delta^t\|_2^2 \|\Delta^{t+1}\|_2. \quad (34)$$

Our final step is to lower bound the left-hand side (LHS) of this inequality. By definition of Z_2 , we have

$$\begin{aligned} \frac{\|S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2}{Z_2^t} &\geq \langle \Delta^{t+1}, \nabla^2 f(x^t) \Delta^{t+1} \rangle \\ &= \langle \Delta^{t+1}, \nabla^2 f(x^*) \Delta^{t+1} \rangle + \langle \Delta^{t+1}, (\nabla^2 f(x^t) - \nabla^2 f(x^*)) \Delta^{t+1} \rangle \\ &\geq \{\gamma - L \|\Delta^t\|_2\} \|\Delta^{t+1}\|_2^2. \end{aligned}$$

Substituting this lower bound into the previous inequality (34) and then rearranging, we find that, as long as $\|\Delta^t\|_2 < \frac{\gamma}{2L}$, we also have $\|\Delta^t\|_2 < \frac{\beta}{2L}$ and consequently

$$\begin{aligned}\|\Delta^{t+1}\|_2 &\leq \frac{2Z_1^t(\beta + L\|\Delta^t\|_2)}{Z_2^t(\gamma - L\|\Delta^t\|_2)}\|\Delta^t\|_2 + \frac{2L}{(\gamma - L\|\Delta^t\|_2)Z_2^t}\|\Delta^t\|_2^2 \\ &\leq \frac{6\beta Z_1^t}{\gamma Z_2^t}\|\Delta^t\|_2 + \frac{4L}{\gamma Z_2^t}\|\Delta^t\|_2^2,\end{aligned}$$

as claimed.

6.2 Proof of Theorem 2

Recall that in this case, we assume that f is a self-concordant strictly convex function. We adopt the following notation and conventions from the book [18]. For a given $x \in \mathbb{R}^d$, we define the pair of dual norms

$$\|u\|_x := \langle \nabla^2 f(x)u, u \rangle^{1/2}, \quad \text{and} \quad \|v\|_x^* := \langle \nabla^2 f(x)^{-1}v, v \rangle^{1/2},$$

as well as the Newton decrement

$$\lambda_f(x) = \langle \nabla^2 f(x)^{-1} \nabla f(x), \nabla f(x) \rangle^{1/2} = \|\nabla^2 f(x)^{-1} \nabla f(x)\|_x = \|\nabla^2 f(x)^{-1/2} \nabla f(x)\|_2.$$

Note that $\nabla^2 f(x)^{-1}$ is well-defined for strictly convex self-concordant functions. In terms of this notation, the exact Newton update is given by $x \mapsto x_{\text{NE}} := x + v$, where

$$v_{\text{NE}} := \arg \min_{z \in \mathcal{C}-x} \underbrace{\left\{ \frac{1}{2} \|\nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle \right\}}_{\Phi(z)}, \quad (35)$$

whereas the Newton sketch update is given by $x \mapsto x_{\text{NSK}} := x + v_{\text{NSK}}$, where

$$v_{\text{NSK}} := \arg \min_{z \in \mathcal{C}-x} \left\{ \frac{1}{2} \|S \nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle \right\}. \quad (36)$$

The proof of Theorem 2 given in this section involves the unconstrained case ($\mathcal{C} = \mathbb{R}^d$), whereas the proofs of later theorems involve the more general constrained case. In the unconstrained case, the two updates take the simpler forms

$$x_{\text{NE}} = x - (\nabla^2 f(x))^{-1} \nabla f(x), \quad \text{and} \quad x_{\text{NSK}} = x - (\nabla^2 f(x)^{1/2} S^T S \nabla^2 f(x)^{1/2})^{-1} \nabla f(x).$$

For a self-concordant function, the sub-optimality of the Newton iterate x_{NE} in function value satisfies the bound

$$f(x_{\text{NE}}) - \underbrace{\min_{x \in \mathbb{R}^d} f(x)}_{f(x^*)} \leq [\lambda_f(x_{\text{NE}})]^2.$$

This classical bound is not directly applicable to the Newton sketch update, since it involves the *approximate* Newton decrement $\lambda_f(x) := -\langle \nabla f(x), v_{\text{NSK}} \rangle$, as opposed to the exact one $\lambda_f(x) := -\langle \nabla f(x), v_{\text{NE}} \rangle$. Thus, our strategy is to prove that with high probability over the randomness in the sketch matrix, the approximate Newton decrement can be used as an exit condition.

Recall the definitions (35) and (36) of the exact v_{NE} and sketched Newton v_{NSK} update directions, as well as the definition of the tangent cone \mathcal{K} at $x \in \mathcal{C}$. Let \mathcal{K}^t be the tangent cone at x^t . The following lemma provides a high probability bound on their difference:

Lemma 2. Let $S \in \mathbb{R}^{m \times n}$ be a sub-Gaussian or ROS sketch matrix, and consider any fixed vector $x \in \mathcal{C}$ independent of the sketch matrix. If $m \geq c_0 \frac{\mathcal{W}(\nabla^2 f(x)^{1/2} \mathcal{K}^t)^2}{\epsilon^2}$, then

$$\left\| \nabla^2 f(x)^{1/2} (v_{NSK} - v_{NE}) \right\|_2 \leq \epsilon \left\| \nabla^2 f(x)^{1/2} v_{NE} \right\|_2 \quad (37)$$

with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

Similar to the standard analysis of Newton's method, our analysis of the Newton sketch algorithm is split into two phases defined by the magnitude of the decrement $\tilde{\lambda}_f(x)$. In particular, the following lemma constitute the core of our proof:

Lemma 3. For $\epsilon \in (0, 1/2)$, there exist constants $\nu > 0$ and $\eta \in (0, 1/16)$ such that:

(a) If $\tilde{\lambda}_f(x) > \eta$, then $f(x_{NSK}) - f(x) \leq -\nu$ with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

(b) Conversely, if $\tilde{\lambda}_f(x) \leq \eta$, then

$$\tilde{\lambda}_f(x_{NSK}) \leq \tilde{\lambda}_f(x), \quad \text{and} \quad (38a)$$

$$\lambda_f(x_{NSK}) \leq \left(\frac{16}{25}\right) \lambda_f(x), \quad (38b)$$

where both bounds hold with probability $1 - c_1 e^{-c_2 m \epsilon^2}$.

Using this lemma, let us now complete the proof of the theorem, dividing our analysis into the two phases of the algorithm.

First phase analysis: By Lemma 3(a) each iteration in the first phase decreases the function value by at least $\nu > 0$, the number of first phase iterations N_1 is at most

$$N_1 := \frac{f(x^0) - f(x^*)}{\nu},$$

with probability at least $1 - N_1 c_1 e^{-c_2 m}$.

Second phase analysis: Next, let us suppose that at some iteration t , the condition $\tilde{\lambda}_f(x^t) \leq \eta$ holds, so that part (b) of Lemma 3 can be applied. In fact, the bound (38a) then guarantees that $\tilde{\lambda}_f(x^{t+1}) \leq \eta$, so that we may apply the contraction bound (38b) repeatedly for N_2 rounds so as to obtain that

$$\lambda_f(x^{t+N_2}) \leq \left(\frac{16}{25}\right)^{N_2} \lambda_f(x^t)$$

with probability $1 - N_2 c_1 e^{-c_2 m}$.

Since $\lambda_f(x^t) \leq \eta \leq 1/16$ by assumption, the self-concordance of f then implies that

$$f(x^{t+k}) - f(x^*) \leq \left(\frac{16}{25}\right)^k \frac{1}{16}.$$

Therefore, in order to ensure that and consequently for achieving $f(x^{t+k}) - f(x^*) \leq \epsilon$, it suffices to the number of second phase iterations lower bounded as $N_2 \geq 0.65 \log_2(\frac{1}{16\epsilon})$.

Putting together the two phases, we conclude that the total number of iterations N required to achieve ϵ -accuracy is at most

$$N = N_1 + N_2 \leq \frac{f(x^0) - f(x^*)}{\gamma} + 0.65 \log_2\left(\frac{1}{16\epsilon}\right),$$

and moreover, this guarantee holds with probability at least $1 - Nc_1e^{-c_2m\epsilon^2}$.

The final step in our proof of the theorem is to establish Lemma 3, and we do in the next two subsections.

6.2.1 Proof of Lemma 3(a)

Our proof of this part is performed conditionally on the event $\mathcal{D} := \{\tilde{\lambda}_f(x) > \eta\}$. Our strategy is to show that the backtracking line search leads to a stepsize $s > 0$ such that function decrement in moving from the current iterate x to the new sketched iterate $x_{\text{NSK}} = x + sv_{\text{NSK}}$ is at least

$$f(x_{\text{NSK}}) - f(x) \leq -\nu \quad \text{with probability at least } 1 - c_1e^{-c_2m}. \quad (39)$$

The outline of our proof is as follows. Defining the univariate function $g(u) := f(x + uv_{\text{NSK}})$ and $\epsilon' = \frac{2\epsilon}{1-\epsilon}$, we first show that $\hat{u} = \frac{1}{1+(1+\epsilon')\tilde{\lambda}_f(x)}$ satisfies the bound

$$g(\hat{u}) \leq g(0) - a\hat{u}\tilde{\lambda}_f(x)^2, \quad (40a)$$

which implies that \hat{u} satisfies the exit condition of backtracking line search. Therefore, the stepsize s must be lower bounded as $s \geq b\hat{u}$, which then implies that the updated solution $x_{\text{NSK}} = x + sv_{\text{NSK}}$ satisfies the decrement bound

$$f(x_{\text{NSK}}) - f(x) \leq -ab \frac{\tilde{\lambda}_f(x)^2}{1 + (1 + \frac{2\epsilon}{1-\epsilon})\tilde{\lambda}_f(x)}. \quad (40b)$$

Since $\tilde{\lambda}_f(x) > \eta$ by assumption and the function $u \rightarrow \frac{u^2}{1+(1+\frac{2\epsilon}{1-\epsilon})u}$ is monotone increasing, this bound implies that inequality (39) holds with $\nu = ab \frac{\eta^2}{1+(1+\frac{2\epsilon}{1-\epsilon})\eta}$.

It remains to prove the claims (40a) and (40b), for which we make use of the following auxiliary lemma:

Lemma 4. *For $u \in \text{dom } g \cap \mathbb{R}^+$, we have the decrement bound*

$$g(u) \leq g(0) + u\langle \nabla f(x), v_{\text{NSK}} \rangle - u\|[\nabla^2 f(x)]^{1/2}v_{\text{NSK}}\|_2 - \log(1 - u\|[\nabla^2 f(x)]^{1/2}v_{\text{NSK}}\|_2). \quad (41)$$

provided that $u\|[\nabla^2 f(x)]^{1/2}v_{\text{NSK}}\|_2 < 1$.

Lemma 5. *With probability at least $1 - c_1e^{-c_2m}$, we have*

$$\|[\nabla^2 f(x)]^{1/2}v_{\text{NSK}}\|_2^2 \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 [\tilde{\lambda}_f(x)]^2. \quad (42)$$

The proof of these lemmas are provided in Appendices A.2 and A.3. Using them, let us prove the claims (40a) and (40b). Recalling our shorthand $\epsilon' := \frac{1+\epsilon}{1-\epsilon} - 1 = \frac{2\epsilon}{1-\epsilon}$, substituting inequality (42) into the decrement formula (41) yields

$$\begin{aligned} g(u) &\leq g(0) - u\tilde{\lambda}_f(x)^2 - u(1+\epsilon') \tilde{\lambda}_f(x) - \log(1 - u(1+\epsilon') \tilde{\lambda}_f(x)) \\ &= g(0) - \left\{ u(1+\epsilon')^2 \tilde{\lambda}_f(x)^2 + u(1+\epsilon') \tilde{\lambda}_f(x) + \log(1 - u(1+\epsilon') \tilde{\lambda}_f(x)) \right\} \\ &\quad + u((1+\epsilon')^2 - 1)\tilde{\lambda}_f(x)^2 \end{aligned} \tag{43}$$

where we added and subtracted $u(1+\epsilon')^2 \tilde{\lambda}_f(x)^2$ so as to obtain the final equality.

We now prove inequality (40a). Now setting $u = \hat{u} := \frac{1}{1+(1+\epsilon')\tilde{\lambda}_f(x)}$, which satisfies the conditions of Lemma 4 yields

$$g(\hat{u}) \leq g(0) - (1+\epsilon') \tilde{\lambda}_f(x) + \log(1 + (1+\epsilon') \tilde{\lambda}_f(x)) + \frac{(\epsilon'^2 + 2\epsilon')\tilde{\lambda}_f(x)^2}{1 + (1+\epsilon')\tilde{\lambda}_f(x)}.$$

Making use of the standard inequality $-u + \log(1+u) \leq -\frac{\frac{1}{2}u^2}{(1+u)}$ (for instance, see the book [4]), we find that

$$\begin{aligned} g(\hat{u}) &\leq g(0) - \frac{\frac{1}{2}(1+\epsilon')^2 \tilde{\lambda}_f(x)^2}{1 + (1+\epsilon')\tilde{\lambda}_f(x)} + \frac{(\epsilon'^2 + 2\epsilon')\tilde{\lambda}_f(x)^2}{1 + (1+\epsilon')\tilde{\lambda}_f(x)} \\ &= g(0) - \left(\frac{1}{2} - \frac{1}{2}\epsilon'^2 - \epsilon' \right) \tilde{\lambda}_f(x)^2 \hat{u} \\ &\leq g(0) - \alpha \tilde{\lambda}_f(x)^2 \hat{u}, \end{aligned}$$

where the final inequality follows from our assumption $\alpha \leq \frac{1}{2} - \frac{1}{2}\epsilon'^2 - \epsilon'$. This completes the proof of the bound (40a). Finally, the lower bound (40b) follows by setting $u = b\hat{u}$ into the decrement inequality (41).

6.2.2 Proof of Lemma 3(b)

The proof of this part hinges on the following auxiliary lemma:

Lemma 6. *For all $\epsilon \in (0, 1/2)$, we have*

$$\lambda_f(x_{NSK}) \leq \frac{(1+\epsilon)\lambda_f^2(x) + \epsilon\lambda_f(x)}{\left(1 - (1+\epsilon)\lambda_f(x)\right)^2}, \quad \text{and} \tag{44a}$$

$$(1-\epsilon)\lambda_f(x_{NSK}) \leq \tilde{\lambda}_f(x_{NSK}) \leq (1+\epsilon)\lambda_f(x_{NSK}), \tag{44b}$$

where all bounds hold with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

See Appendix A.4 for the proof.

We now use Lemma 6 to prove the two claims in the lemma statement.

Proof of the bound (38a): Recall from the theorem statement that $\eta := \frac{1}{8} \frac{1 - \frac{1}{2}(\frac{1+\epsilon}{1-\epsilon})^2 - a}{(\frac{1+\epsilon}{1-\epsilon})^3}$. By examining the roots of a polynomial in ϵ , it can be seen that $\eta \leq \frac{1-\epsilon}{1+\epsilon} \frac{1}{16}$.

$$(1 + \epsilon) \lambda_f(x^t) \leq (1 + \epsilon) \tilde{\lambda}_f(x^t) \leq (1 + \epsilon) \eta \leq \frac{1}{16}$$

By applying the inequalities (44b), we have

$$(1 + \epsilon) \lambda_f(x) \leq \frac{1 + \epsilon}{1 - \epsilon} \tilde{\lambda}_f(x) \leq \frac{1 + \epsilon}{1 - \epsilon} \eta \leq \frac{1}{16} \quad (45)$$

whence inequality (44a) implies that

$$\lambda_f(x_{\text{NSK}}) \leq \frac{\frac{1}{16} \lambda_f(x) + \epsilon \lambda_f(x)}{(1 - \frac{1}{16})^2} \leq \left(\frac{16}{225} + \frac{256}{225} \epsilon \right) \lambda_f(x) \leq \frac{16}{25} \lambda_f(x). \quad (46)$$

Here the final inequality holds for all $\epsilon \in (0, 1/2)$. Combining the bound (44b) with inequality (46) yields

$$\tilde{\lambda}_f(x_{\text{NSK}}) \leq (1 + \epsilon) \lambda_f(x_{\text{NSK}}) \leq (1 + \epsilon) \left(\frac{16}{25} \right) \tilde{\lambda}_f(x) \leq \tilde{\lambda}_f(x),$$

where the final inequality again uses the condition $\epsilon \in (0, \frac{1}{2})$. This completes the proof of the bound (38a).

Proof of the bound (38b): This inequality has been established as a consequence of proving the bound (46).

6.3 Proof of Theorem 3

Given the proof of Theorem 2, it remains only to prove the following modified version of Lemma 2. It applies to the exact and sketched Newton directions $v_{\text{NE}}, v_{\text{NSK}} \in \mathbb{R}^d$ that are defined as follows

$$v_{\text{NE}} := \arg \min_{z \in \mathcal{C} - x} \left\{ \frac{1}{2} \|\nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle + \frac{1}{2} \langle z, \nabla^2 g(x) z \rangle \right\}, \quad (47a)$$

$$v_{\text{NSK}} = \arg \min_{z \in \mathcal{C} - x} \underbrace{\left\{ \frac{1}{2} \|S \nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle + \frac{1}{2} \langle z, \nabla^2 g(x) z \rangle \right\}}_{\Psi(z; S)}. \quad (47b)$$

Thus, the only difference is that the Hessian $\nabla^2 f(x)$ is sketched, whereas the term $\nabla^2 g(x)$ remains unsketched.

Lemma 7. *Let $S \in \mathbb{R}^{m \times n}$ be a sub-Gaussian or ROS sketching matrix, and let $x \in \mathbb{R}^d$ be a (possibly random) vector independent of S . If $m \geq c_0 \max_{x \in \mathcal{C}} \frac{\mathcal{W}(\nabla^2 f(x)^{1/2} \mathcal{K})^2}{\epsilon^2}$, then*

$$\left\| \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \leq \epsilon \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 \quad (48)$$

with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

7 Discussion

In this paper, we introduced and analyzed the Newton sketch, a randomized approximation to the classical Newton updates. This algorithm is a natural generalization of the Iterative Hessian Sketch (IHS) updates analyzed in our earlier work [19]. The IHS applies only to constrained least-squares problems (for which the Hessian is independent of the iteration number), whereas the Newton Sketch applies to any any twice differentiable function subject to a closed convex constraint set. We described various applications of the Newton sketch, including its use with barrier methods to solve various forms of constrained problems. For the minimization of self-concordant functions, the combination of the Newton sketch within interior point updates leads to much faster algorithms for an extensive body of convex optimization problems.

Each iteration of the Newton sketch always has lower computational complexity than classical Newton’s method. Moreover, it has lower computational complexity than first-order methods when either $n \geq d^2$ or $d \geq n^2$ (using the dual strategy); here n and d denote the dimensions of the data matrix A . In the context of barrier methods, the parameters n and d typically correspond to the number of constraints and number of variables, respectively. In many “big data” problems, one of the dimensions is much larger than the other, in which case the Newton sketch is advantageous. Moreover, sketches based on the randomized Hadamard transform are well-suited to in parallel environments: in this case, the sketching step can be done in $\mathcal{O}(\log m)$ time with $\mathcal{O}(nd)$ processors. This scheme significantly decreases the amount of central computation—namely, from $\mathcal{O}(m^2d + nd \log m)$ to $\mathcal{O}(m^2d + \log d)$.

There are a number of open problems associated with the Newton sketch. Here we focused our analysis on the cases of sub-Gaussian and randomized orthogonal system (ROS) sketches. It would also be interesting to analyze sketches based on coordinate sampling, or other forms of “sparse” sketches (for instance, see the paper [10]). Such techniques might lead to significant gains in cases where the data matrix A is itself sparse: more specifically, it may be possible to obtain sketched optimization algorithms whose computational complexity only scales with number of nonzero entries in the data matrices the full dimensionality nd . Finally, it would be interesting to explore the problem of lower bounds on the sketch dimension m . In particular, is there a threshold below which any algorithm that has access only to gradients and m -sketched Hessians must necessarily converge at a sub-linear rate, or in a way that depends on the strong convexity and smoothness parameters? Such a result would clarify whether or not the guarantees in this paper are improvable.

Acknowledgements

Both authors were partially supported by Office of Naval Research MURI grant N00014-11-1-0688, and National Science Foundation Grants CIF-31712-23800 and DMS-1107000. In addition, MP was supported by a Microsoft Research Fellowship.

A Technical results for Theorem 2

In this appendix, we collect together various technical results and proofs that are required in the proof of Theorem 2.

A.1 Proof of Lemma 2

Let u be a unit-norm vector independent of S , and consider the random quantities

$$Z_1(S, x) := \inf_{v \in \nabla^2 f(x)^{1/2} \mathcal{K}^t \cap \mathcal{S}^{n-1}} \|Sv\|_2^2 \quad \text{and} \quad (49a)$$

$$Z_2(S, x) := \sup_{v \in \nabla^2 f(x)^{1/2} \mathcal{K}^t \cap \mathcal{S}^{n-1}} \left| \langle u, (S^T S - I_n) v \rangle \right|. \quad (49b)$$

By the optimality and feasibility of v_{NSK} and v_{NE} (respectively) for the sketched Newton update (36), we have

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} v_{\text{NSK}}\|_2^2 - \langle v_{\text{NSK}}, \nabla f(x) \rangle \leq \frac{1}{2} \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2^2 - \langle v_{\text{NE}}, \nabla f(x) \rangle.$$

Defining the difference vector $\hat{e} := v_{\text{NSK}} - v_{\text{NE}}$, some algebra leads to the basic inequality

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 \leq -\langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, S^T S \nabla^2 f(x)^{1/2} \hat{e} \rangle + \langle \hat{e}, \nabla f(x) \rangle. \quad (50)$$

Moreover, by the optimality and feasibility of v_{NE} and v_{NSK} for the exact Newton update (35), we have

$$\langle \nabla^2 f(x) v_{\text{NE}} - \nabla f(x), \hat{e} \rangle = \langle \nabla^2 f(x) v_{\text{NE}} - \nabla f(x), v_{\text{NSK}} - v_{\text{NE}} \rangle \geq 0. \quad (51)$$

Consequently, by adding and subtracting $\langle \nabla^2 f(x) v_{\text{NE}}, \hat{e} \rangle$, we find that

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 \leq \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right|. \quad (52)$$

By definition, the error vector \hat{e} belongs to the cone \mathcal{K}^t and the vector $\nabla^2 f(x)^{1/2} v_{\text{NE}}$ is fixed and independent of the sketch. Consequently, invoking definitions (49a) and (49b) of the random variables Z_1 and Z_2 yields

$$\begin{aligned} \frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 &\geq \frac{Z_1}{2} \|\nabla^2 f(x)^{1/2} \hat{e}\|_2^2, \\ \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right| &\leq Z_2 \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 \|\nabla^2 f(x)^{1/2} \hat{e}\|_2, \end{aligned}$$

Putting together the pieces, we find that

$$\left\| \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \leq \frac{2Z_2(S, x)}{Z_1(S, x)} \left\| \nabla^2 f(x)^{1/2} (v_{\text{NE}}) \right\|_2. \quad (53)$$

Finally, for any $\delta \in (0, 1)$, let us define the event $\mathcal{E}(\delta) = \{Z_1 \geq 1 - \delta, \text{ and } Z_2 \leq \delta\}$. By Lemma 4 and Lemma 5 from our previous paper [20], we are guaranteed that $\mathbb{P}[\mathcal{E}(\delta)] \geq 1 - c_1 e^{-c_2 m \delta^2}$. Conditioned on the event $\mathcal{E}(\delta)$, the bound (53) implies that

$$\left\| \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \leq \frac{2\delta}{1 - \delta} \left\| \nabla^2 f(x)^{1/2} (v_{\text{NE}}) \right\|_2.$$

By setting $\delta = \frac{\epsilon}{4}$, the claim follows.

A.2 Proof of Lemma 4

By construction, the function $g(u) = f(x + uv_{\text{NSK}})$ is strictly convex and self-concordant. Consequently, it satisfies the bound $\frac{d}{du} (g''(u)^{-1/2}) \leq 1$, whence

$$g''(s)^{-1/2} - g''(0)^{-1/2} = \int_0^s \frac{d}{du} (g''(u)^{-1/2}) du \leq s.$$

or equivalently $g''(s) \leq \frac{g''(0)}{(1 - sg''(0)^{1/2})^2}$ for $s \in \text{dom } g \cap [0, g''(0)^{-1/2})$. Integrating this inequality twice yields the bound

$$g(u) \leq g(0) + ug'(0) - ug''(0)^{1/2} - \log(1 - ug''(0)^{1/2}). \quad (54)$$

Since $g'(u) = \langle \nabla f(x + uv_{\text{NSK}}), v_{\text{NSK}} \rangle$ and $g''(u) = \langle v_{\text{NSK}}, \nabla^2 f(x + uv_{\text{NSK}}) v_{\text{NSK}} \rangle$, the decrement bound (41) follows.

A.3 Proof of Lemma 5

We perform this analysis conditional on the bound (37) from Lemma 2. We begin by observing that

$$\begin{aligned} \|[\nabla^2 f(x)]^{1/2} v_{\text{NSK}}\|_2 &\leq \|[\nabla^2 f(x)]^{1/2} v_{\text{NE}}\|_2 + \|[\nabla^2 f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &= \lambda_f(x) + \|[\nabla^2 f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2. \end{aligned} \quad (55)$$

Lemma 2 implies that $\|[\nabla^2 f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \epsilon \|[\nabla^2 f(x)]^{1/2} v_{\text{NE}}\|_2 = \epsilon \lambda_f(x)$. In conjunction with the bound (56), we see that

$$\|[\nabla^2 f(x)]^{1/2} v_{\text{NSK}}\|_2 \leq (1 + \epsilon) \lambda_f(x). \quad (56)$$

Our next step is to lower bound the term $\langle \nabla f(x), v_{\text{NSK}} \rangle$: in particular, by adding and subtracting a factor of the original Newton step v_{NE} , we find that

$$\begin{aligned} \langle \nabla f(x), v_{\text{NSK}} \rangle &= \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2[f(x)]^{1/2} v_{\text{NSK}} \rangle \\ &= \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2[f(x)]^{1/2} v_{\text{NE}} \rangle + \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2[f(x)]^{1/2} (v_{\text{NSK}} - v) \rangle \\ &= -\|[\nabla^2 f(x)]^{-1/2} \nabla f(x)\|_2^2 + \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2[f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \rangle \\ &\leq -\|[\nabla^2 f(x)]^{-1/2} \nabla f(x)\|_2^2 + \|[\nabla^2 f(x)]^{-1/2} \nabla f(x)\|_2 \|\nabla^2[f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &= -\lambda_f(x)^2 + \lambda_f(x) \|\nabla^2[f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &\leq -\lambda_f(x)^2 (1 - \epsilon), \end{aligned} \quad (57)$$

where the final step again makes use of Lemma 2. Repeating the above argument in the reverse direction yields the lower bound $\langle \nabla f(x), v_{\text{NSK}} \rangle \geq -\lambda_f(x)^2 (1 + \epsilon)$, so that we may conclude that

$$|\tilde{\lambda}_f(x^t) - \lambda_f(x^t)| \leq \epsilon \lambda_f(x^t). \quad (58)$$

Finally, by squaring both sides of the inequality (56) and combining with the above bounds gives

$$\|[\nabla^2 f(x)]^{1/2} v_{\text{NSK}}\|_2^2 \leq \frac{(1 + \epsilon)^2}{1 - \epsilon} \langle \nabla f(x), v_{\text{NSK}} \rangle = \frac{(1 + \epsilon)^2}{1 - \epsilon} \tilde{\lambda}_f^2(x) \leq \left(\frac{1 + \epsilon}{1 - \epsilon} \right)^2 \tilde{\lambda}_f^2(x),$$

as claimed.

A.4 Proof of Lemma 6

We have already proved the bound (44b) during our proof of Lemma 5—in particular, see equation (58). Accordingly, it remains only to prove the inequality (44a).

Introducing the shorthand $\tilde{\lambda} := (1 + \epsilon)\lambda_f(x)$, we first claim that the Hessian satisfies the sandwich relation

$$(1 - s\alpha)^2 \nabla^2 f(x) \preceq \nabla^2 f(x + sv_{\text{NSK}}) \preceq \frac{1}{(1 - s\alpha)^2} \nabla^2 f(x), \quad (59)$$

for $|1 - s\alpha| < 1$ where $\alpha = (1 + \epsilon)\lambda_f(x)$, with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$. Let us recall Theorem 4.1.6 of Nesterov [17]: it guarantees that

$$(1 - s\|v_{\text{NSK}}\|_x)^2 \nabla^2 f(x) \preceq \nabla^2 f(x + sv_{\text{NSK}}) \preceq \frac{1}{(1 - s\|v_{\text{NSK}}\|_x)^2} \nabla^2 f(x). \quad (60)$$

Now recall the bound (37) from Lemma 2: combining it with an application of the triangle inequality (in terms of the semi-norm $\|v\|_x = \|\nabla^2 f(x)^{1/2} v\|_2$) yields

$$\left\| \nabla^2 f(x)^{1/2} v_{\text{NSK}} \right\|_2 \leq (1 + \epsilon) \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 = (1 + \epsilon) \|v_{\text{NE}}\|_x,$$

with probability at least $1 - e^{-c_1 m \epsilon^2}$, and substituting this inequality into the bound (60) yields the sandwich relation (59) for the Hessian.

Using this sandwich relation (59), the Newton decrement can be bounded as

$$\begin{aligned} \lambda_f(x_{\text{NSK}}) &= \|\nabla^2 f(x_{\text{NSK}})^{-1/2} \nabla f(x_{\text{NSK}})\|_2 \\ &\leq \frac{1}{(1 - (1 + \epsilon)\lambda_f(x))} \|\nabla^2 f(x)^{-1/2} \nabla f(x_{\text{NSK}})\|_2 \\ &= \frac{1}{(1 - (1 + \epsilon)\lambda_f(x))} \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + sv_{\text{NSK}}) v_{\text{NSK}} ds \right) \right\|_2 \\ &= \frac{1}{(1 - (1 + \epsilon)\lambda_f(x))} \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + sv_{\text{NSK}}) v_{\text{NE}} ds + \Delta \right) \right\|_2, \end{aligned}$$

where we have defined $\Delta = \int_0^1 \nabla^2 f(x + sv_{\text{NSK}}) (v_{\text{NSK}} - v_{\text{NE}}) ds$. By the triangle inequality, we can write $\lambda_f(x_{\text{NSK}}) \leq \frac{1}{(1 - (1 + \epsilon)\lambda_f(x))} (M_1 + M_2)$, where

$$M_1 := \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + tv_{\text{NSK}}) v_{\text{NE}} dt \right) \right\|_2, \quad \text{and} \quad M_2 := \left\| \nabla^2 f(x)^{-1/2} \Delta \right\|_2.$$

In order to complete the proof, it suffices to show that

$$M_1 \leq \frac{(1 + \epsilon)\lambda_f(x)^2}{1 - (1 + \epsilon)\lambda_f(x)}, \quad \text{and} \quad M_2 \leq \frac{\epsilon\lambda_f(x)}{1 - (1 + \epsilon)\lambda_f(x)}.$$

Bound on M_1 : Re-arranging and then invoking the Hessian sandwich relation (59) yields

$$\begin{aligned}
M_1 &= \left\| \int_0^1 \left(\nabla^2 f(x)^{-1/2} \nabla^2 f(x + s v_{\text{NSK}}) \nabla^2 f(x)^{-1/2} - I \right) ds \left(\nabla^2 f(x)^{1/2} v_{\text{NE}} \right) \right\|_2 \\
&\leq \left| \int_0^1 \left(\frac{1}{(1 - s(1 + \epsilon)\lambda_f(x))^2} - 1 \right) ds \right| \left\| \left(\nabla^2 f(x)^{1/2} v_{\text{NE}} \right) \right\|_2 \\
&= \frac{(1 + \epsilon)\lambda_f(x)}{1 - (1 + \epsilon)\lambda_f(x)} \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 \\
&= \frac{(1 + \epsilon)\lambda_f^2(x)}{1 - (1 + \epsilon)\lambda_f(x)}.
\end{aligned}$$

Bound on M_2 : We have

$$\begin{aligned}
M_2 &= \left\| \int_0^1 \nabla^2 f(x)^{-1/2} \nabla^2 f(x + s v_{\text{NSK}}) \nabla^2 f(x)^{-1/2} ds \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\
&\leq \left\| \int_0^1 \frac{1}{(1 - s(1 + \epsilon)\lambda_f(x))^2} ds \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\
&= \frac{1}{1 - (1 + \epsilon)\lambda_f(x)} \left\| \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\
&\stackrel{(i)}{\leq} \frac{1}{1 - (1 + \epsilon)\lambda_f(x)} \epsilon \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 \\
&= \frac{\epsilon \lambda_f(x)}{1 - (1 + \epsilon)\lambda_f(x)},
\end{aligned}$$

where the inequality in step (i) follows from Lemma 2.

B Proof of Lemma 7

The proof follows the basic inequality argument of the proof of Lemma 2. Since v_{NSK} and v_{NE} are optimal and feasible (respectively) for the sketched Newton problem (47b), we have $\Psi(v_{\text{NSK}}; S) \leq \Psi(v_{\text{NE}}; S)$. Defining the difference vector $\hat{e} := v_{\text{NSK}} - v$, some algebra leads to the basic inequality

$$\begin{aligned}
\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 + \frac{1}{2} \langle \hat{e}, \nabla^2 g(x) \hat{e} \rangle &\leq -\langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, S^T S \nabla^2 f(x)^{1/2} \hat{e} \rangle \\
&\quad + \langle \hat{e}, (\nabla f(x) - \nabla^2 g(x)) v_{\text{NE}} \rangle.
\end{aligned}$$

On the other hand since v_{NE} and v_{NSK} are optimal and feasible (respectively) for the Newton step (47a), we have

$$\langle \nabla^2 f(x) v_{\text{NE}} + \nabla^2 g(x) v_{\text{NE}} - \nabla f(x), \hat{e} \rangle \geq 0.$$

Consequently, by adding and subtracting $\langle \nabla^2 f(x) v_{\text{NE}}, \hat{e} \rangle$, we find that

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 + \frac{1}{2} \langle v_{\text{NE}}, \nabla^2 g(x) v_{\text{NE}} \rangle \leq \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right|.$$

By convexity of g , we have $\nabla^2 g(x) \succeq 0$, whence

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 \leq \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right|$$

Given this inequality, the remainder of the proof follows as in the proof of Lemma 2.

C Gaussian widths with ℓ_1 -constraints

In this appendix, we state and prove an elementary lemma that bounds for the Gaussian width for a broad class of ℓ_1 -constrained problems. In particular, given a twice-differentiable convex function ψ , a vector $c \in \mathbb{R}^d$, a radius R and a collection of d -vectors $\{a_i\}_{i=1}^n$, consider a convex program of the form

$$\min_{x \in \mathcal{C}} \left\{ \sum_{i=1}^n \psi(\langle a_i, x \rangle) + \langle c, x \rangle \right\}, \quad \text{where } \mathcal{C} = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq R\}. \quad (61)$$

Lemma 8. *Suppose that the ℓ_1 -constrained program (61) has a unique optimal solution x^* such that $\|x^*\|_0 \leq s$ for some integer s . Then denoting the tangent cone at x^* by \mathcal{K} , then*

$$\max_{x \in \mathcal{C}} \mathcal{W}(\nabla^2 f(x)^{1/2} \mathcal{K}) \leq 6\sqrt{s \log d} \sqrt{\frac{\psi''_{\max}}{\psi''_{\min}}} \frac{\max_{j=1,\dots,d} \|A_j\|_2}{\sqrt{\gamma_s^-(A)}},$$

where

$$\psi''_{\min} = \min_{x \in \mathcal{C}} \min_{i=1,\dots,n} \psi''(\langle a_i, x \rangle, y_i), \quad \text{and} \quad \psi''_{\max} = \max_{x \in \mathcal{C}} \max_{i=1,\dots,n} \psi''(\langle a_i, x \rangle, y_i).$$

Proof. It is well known [16, 20] that the tangent cone of the ℓ_1 -norm at any s -sparse solution is a subset of the cone $\{z \in \mathbb{R}^d \mid \|z\|_1 \leq 2\sqrt{s}\|z\|_2\}$. Using this fact, we have the following sequence of upper bounds

$$\begin{aligned} \mathcal{W}(\nabla^2 f(x)^{1/2} \mathcal{K}) &= \mathbb{E}_w \max_{\substack{z^T \nabla^2 f(x) z = 1, \\ z \in \mathcal{K}}} \langle w, \nabla^2 f(x)^{1/2} z \rangle \\ &= \mathbb{E}_w \max_{\substack{z^T A^T \text{diag}(\psi''(\langle a_i, x \rangle, y_i)) A z = 1, \\ z \in \mathcal{K}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &\leq \mathbb{E}_w \max_{\substack{z^T A^T A z \leq 1/\psi''_{\min}, \\ z \in \mathcal{K}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &\leq \mathbb{E}_w \max_{\|z\|_1 \leq \frac{2\sqrt{s}}{\sqrt{\gamma_s^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &= \frac{2\sqrt{s}}{\sqrt{\gamma_s^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}} \mathbb{E}_w \|A^T \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} w\|_{\infty} \\ &= \frac{2\sqrt{s}}{\sqrt{\gamma_s^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}} \mathbb{E}_w \max_{j=1,\dots,d} \left| \underbrace{\sum_{i=1,\dots,n} w_i A_{ij} \psi''(\langle a_i, x \rangle, y_i)^{1/2}}_{Q_j} \right|. \end{aligned}$$

Here the random variables Q_j are zero-mean Gaussians with variance at most

$$\sum_{i=1,\dots,n} A_{ij}^2 \psi''(\langle a_i, x \rangle, y_i) \leq \psi''_{\max} \|A_j\|_2^2.$$

Consequently, applying standard bounds on the suprema of Gaussian variates [12], we obtain

$$\mathbb{E}_w \max_{j=1,\dots,d} \left| \sum_{i=1,\dots,n} w_i A_{ij} \psi''(\langle a_i, x \rangle, y_i)^{1/2} \right| \leq 3\sqrt{\log d} \sqrt{\psi''_{\max}} \max_{j=1,\dots,d} \|A_j\|_2.$$

When combined with the previous inequality, the claim follows. \square

References

- [1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.
- [2] N. Ailon and E. Liberty. Fast dimension reduction using rademacher series on dual bch codes. *Discrete Comput. Geom*, 42(4):615–630, 2009.
- [3] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.
- [4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.
- [5] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdam, NL, 2001.
- [6] P. Drineas, M. Magdon-Ismail, M.W. Mahoney, and D.P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.
- [7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
- [8] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.
- [9] T. Hastie and B. Efron. lars: Least angle regression, lasso and forward stagewise. *R package version 0.9-7*, 2007.
- [10] D.M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4, 2014.
- [11] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1 -regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, 1(4):606–617, 2007.
- [12] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.
- [13] H. M. Markowitz. *Portfolio Selection*. Wiley, New York, 1959.
- [14] P. McCullagh and J.A. Nelder. *Generalized linear models*. Monographs on statistics and applied probability 37. Chapman and Hall/CRC, New York, 1989.
- [15] M. Fukushima N. Yamashita. On the rate of convergence of the levenberg-marquardt method. In *Topics in numerical analysis*, pages 239–249. Springer, 2001.
- [16] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

- [17] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, New York, 2004.
- [18] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics, 1994.
- [19] M. Pilanci and M. J. Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. Technical report, UC Berkeley, 2014. Full length version at arXiv:1411.0347.
- [20] M. Pilanci and M. J. Wainwright. Randomized sketches of convex programs with sharp guarantees. Technical report, UC Berkeley, 2014. Full length version at arXiv:1404.7203; Presented in part at ISIT 2014.
- [21] G. Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and Analysis*, volume 1206 of *Lecture Notes in Mathematics*, pages 167–241. Springer, 1989.
- [22] D.A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.
- [23] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [24] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 2012.
- [25] S.J. Wright and J. Nocedal. *Numerical optimization*, volume 2. Springer New York, 1999.