Geometric Median in Nearly Linear Time

Michael B. Cohen
MIT
micohen@mit.edu

Yin Tat Lee
MIT
yintat@mit.edu

Gary Miller
Carnegie Mellon University
glmiller@cs.cmu.edu

Jakub Pachocki
Carnegie Mellon University
pachocki@cs.cmu.edu

Aaron Sidford
Microsoft Research New England
asid@microsoft.com

Abstract

In this paper we provide faster algorithms for solving the geometric median problem: given n points in $\mathbb{R}^d$ compute a point that minimizes the sum of Euclidean distances to the points. This is one of the oldest non-trivial problems in computational geometry yet despite an abundance of research the previous fastest algorithms for computing a $(1+\epsilon)$-approximate geometric median were $O(d \cdot n^{4/3}\epsilon^{-8/3})$ by Chin et. al, $\tilde{O}(d \exp \epsilon^{-4} \log \epsilon^{-1})$ by Badoiu et. al, $O(nd + \mathrm{poly}(d, \epsilon^{-1})$ by Feldman and Langberg, and $O((nd)^{O(1)} \log \frac{1}{\epsilon})$ by Parrilo and Sturmfels and Xue and Ye.

In this paper we show how to compute a $(1+\epsilon)$-approximate geometric median in time $O(nd \log^3 \frac{1}{\epsilon})$ and $O(d\epsilon^{-2})$. While our $O(d\epsilon^{-2})$ is a fairly straightforward application of stochastic subgradient descent, our $O(nd \log^3 \frac{1}{\epsilon})$ time algorithm is a novel long step interior point method. To achieve this running time we start with a simple $O((nd)^{O(1)} \log \frac{1}{\epsilon})$ time interior point method and show how to improve it, ultimately building an algorithm that is quite non-standard from the perspective of interior point literature. Our result is one of very few cases we are aware of outperforming traditional interior point theory and the only we are aware of using interior point methods to obtain a nearly linear time algorithm for a canonical optimization problem that traditionally requires superlinear time. We hope our work leads to further improvements in this line of research.

1 Introduction

One of the oldest easily-stated nontrivial problems in computational geometry is the Fermat-Weber problem: given a set of n points in d dimensions $a^{(1)}, \ldots, a^{(n)} \in \mathbb{R}^d$, find a point $x_* \in \mathbb{R}^d$ that minimizes the sum of Euclidean distances to them:

$$x_* \in \arg\min_{x \in \mathbb{R}^d} f(x) \quad \text{where} \quad f(x) \stackrel{\text{def}}{=} \sum_{i \in [n]} \|x - a^{(i)}\|_2$$

This problem, also known as the *geometric median problem,* is well studied and has numerous applications. It is often considered over low dimensional spaces in the context of the facility location problem [29] and over higher dimensional spaces it has applications to clustering in machine learning and data analysis. For example, computing the geometric median is a subroutine in popular expectation maximization heuristics for k-medians clustering.

The problem is also important to robust estimation, where we like to find a point representative of given set of points that is resistant to outliers. The geometric median is a rotation and translation invariant estimator that achieves the optimal *breakdown point* of 0.5, i.e. it is a good estimator even when up to half of the input data is arbitrarily corrupted [18]. Moreover, if a large constant fraction of the points lie in a ball of diameter ϵ then the geometric median lies in that ball with diameter $O(\epsilon)$ (see Lemma 24). Consequently, the geometric median can be used to turn expected results into high probability results: e.g. if the $a^{(i)}$ are drawn independently such that $\mathbb{E}\|x - a^{(i)}\|_2 \leq \epsilon$ for some $\epsilon > 0$ and $x \in \mathbb{R}^d$ then this fact, Markov bound, and Chernoff Bound, imply $\|x_* - x\|_2 = O(\epsilon)$ with high probability in n.

Despite the ancient nature of the Fermat-Weber problem and its many uses there are relatively few theoretical guarantees for solving it (see Table 1). To compute a $(1+\epsilon)$-approximate solution, i.e. $x \in \mathbb{R}^d$ with $f(x) \leq (1+\epsilon)f(x_*)$, the previous fastest running times were either $O(d \cdot n^{4/3}\epsilon^{-8/3})$ by [7], $\tilde{O}(d \exp \epsilon^{-4} \log \epsilon^{-1})$ by [1], $\tilde{O}(nd + \text{poly}(d, \epsilon^{-1}))$ by [10], or $O((nd)^{O(1)} \log \frac{1}{\epsilon})$ time by [24, 31]. In this paper we improve upon these running times by providing an $O(nd \log^3 \frac{n}{\epsilon})$ time algorithm[1] as well as an $O(d/\epsilon^2)$ time algorithm, provided we have an oracle for sampling a random $a^{(i)}$. Picking the faster algorithm for the particular value of ϵ improves the running time to $O(nd \log^3 \frac{1}{\epsilon})$. We also extend these results to compute a $(1+\epsilon)$-approximate solution to the more general Weber's problem, $\min_{x \in \mathbb{R}^d} \sum_{i \in [n]} w_i \|x - a^{(i)}\|_2$ for non-negative w_i, in time $O(nd \log^3 \frac{1}{\epsilon})$ (see Appendix F).

Our $O(nd \log^3 \frac{n}{\epsilon})$ time algorithm is a careful modification of standard interior point methods for solving the geometric median problem. We provide a long step interior point method tailored to the geometric median problem for which we can implement every iteration in nearly linear time. While our analysis starts with a simple $O((nd)^{O(1)} \log \frac{1}{\epsilon})$ time interior point method and shows how to improve it, our final algorithm is quite non-standard from the perspective of interior point literature. Our result is one of very few cases we are aware of outperforming traditional interior point theory [20, 17] and the only we are aware of using interior point methods to obtain a nearly linear time algorithm for a canonical optimization problem that traditionally requires superlinear time. We hope our work leads to further improvements in this line of research.

Our $O(d\epsilon^{-2})$ algorithm is a relatively straightforward application of sampling techniques and stochastic subgradient descent. Some additional insight is required simply to provide a rigorous analysis of the robustness of the geometric median and use this to streamline our application of stochastic subgradient descent. We include it for completeness however, we defer its proof to Appendix C. The bulk of the work in this paper is focused on developing our $O(nd \log^3 \frac{n}{\epsilon})$ time algorithm which we believe uses a set of techniques of independent interest.

[1] If z is the total number of nonzero entries in the coordinates of the $a^{(i)}$ then a careful analysis of our algorithm improves our running time to $O(z \log^3 \frac{n}{\epsilon})$.

Year	Authors	Runtime	Comments
1659	Torricelli [28]	-	Assuming $n = 3$
1937	Weiszfeld [30]	-	Does not always converge
1990	Chandrasekaran and Tamir[6]	$\widetilde{O}(n \cdot \text{poly}(d) \log \epsilon^{-1})$	Ellipsoid method
1997	Xue and Ye [31]	$\widetilde{O}(\left(d^3 + d^2 n\right) \sqrt{n} \log \epsilon^{-1})$	Interior point with barrier method
2000	Indyk [13]	$\tilde{O}(dn \cdot \epsilon^{-2})$	Optimizes only over x in the input
2001	Parrilo and Sturmfels [24]	$\widetilde{O}(\text{poly}(n, d) \log \epsilon^{-1})$	Reduction to SDP
2002	Badoiu et al. [1]	$\widetilde{O}(d \cdot \exp(O(\epsilon^{-4})))$	Sampling
2003	Bose et al. [4]	$\widetilde{O}(n)$	Assuming $d, \epsilon^{-1} = O(1)$
2005	Har-Peled and Kushal [12]	$\widetilde{O}(n + \text{poly}(\epsilon^{-1}))$	Assuming $d = O(1)$
2011	Feldman and Langberg [10]	$\widetilde{O}(nd + \text{poly}(d, \epsilon^{-1}))$	Coreset
2013	Chin et al. [7]	$\widetilde{O}(dn^{4/3} \cdot \epsilon^{-8/3})$	Multiplicative weights
-	**This paper**	$O(nd \log^3(n/\epsilon))$	Interior point with custom analysis
-	**This paper**	$O(d\epsilon^{-2})$	Stochastic gradient descent

Table 1: Selected Previous Results.

1.1 Previous Work

The geometric median problem was first formulated for the case of three points in the early 1600s by Pierre de Fermat [14, 9]. A simple elegant ruler and compass construction was given in the same century by Evangelista Torricelli. Such a construction does not generalize when a larger number of points is considered: Bajaj has shown the even for five points, the geometric median is not expressible by radicals over the rationals [2]. Hence, the $(1 + \epsilon)$-approximate problem has been studied for larger values of n.

Many authors have proposed algorithms with runtime polynomial in n, d and $1/\epsilon$. The most cited and used algorithm is Weiszfeld's 1937 algorithm [30]. Unfortunately Weiszfeld's algorithm may not converge and if it does it may do so very slowly. There have been many proposed modifications to Weiszfeld's algorithm [8, 25, 23, 3, 27, 16] that generally give non-asymptotic runtime guarantees. In light of more modern multiplicative weights methods his algorithm can be viewed as a re-weighted least squares iteration. Chin et al. [7] considered the more general L_2 embedding problem: placing the vertices of a graph into $\mathbb{R}^d$, where some of the vertices have fixed positions while the remaining vertices are allowed to float, with the objective of minimizing the sum of the Euclidean edge lengths. Using the multiplicative weights method, they obtained a run time of $O(d \cdot n^{4/3} \epsilon^{-8/3})$ for a broad class of problems, including the geometric median problem.[2]

Many authors consider problems that generalize the Fermat-Weber problem, and obtain algorithms for finding the geometric median as a specialization. Badoiu et al. gave an approximate k-median algorithm by sub-sampling with the runtime for $k = 1$ of $\widetilde{O}(d \cdot \exp(O(\epsilon^{-4})))$ [1]. Parrilo and Sturmfels demonstrated that the problem can be reduced to semidefinite programming, thus obtaining a runtime of $\widetilde{O}(\text{poly}(n, d) \log \epsilon^{-1})$ [24]. Furthermore, Bose et al. gave a linear time algorithm for fixed d and ϵ^{-1}, based on low-dimensional data structures [4] and it has been show how to obtain running times of $\widetilde{O}(nd + \text{poly}(d, \epsilon^{-1}))$ for this problem and a more general class of problems.[12, 10].

An approach very related to ours was studied by Xue and Ye [31]. They give an interior point method with barrier analysis that runs in time $\tilde{O}((d^3 + d^2 n)\sqrt{n} \log \epsilon^{-1})$.

[2] The result of [7] was stated in more general terms than given here. However, it easy to formulate the geometric median problem in their model.

1.2 Overview of $O(nd\log^3 \frac{n}{\epsilon})$ Time Algorithm

Interior Point Primer

Our algorithm is broadly inspired by interior point methods, a broad class of methods for efficiently solving convex optimization problems [32, 22]. Given an instance of the geometric median problem we first put the problem in a more natural form for applying interior point methods. Rather than writing the problem as minimizing a convex function over $\mathbb{R}^d$

$$\min_{x\in\mathbb{R}^d} f(x) \quad \text{where} \quad f(x) \stackrel{\text{def}}{=} \sum_{i\in[n]} \|x - a^{(i)}\|_2 \tag{1.1}$$

we instead write the problem as minimizing a linear function over a convex set:

$$\min_{\{\alpha,x\}\in S} 1^\top \alpha \quad \text{where } S = \left\{\alpha \in \mathbb{R}^n, x \in \mathbb{R}^d \;\middle|\; \|x^{(i)} - a^{(i)}\|_2 \leq \alpha_i \text{ for all } i \in [n]\right\} . \tag{1.2}$$

Clearly, these problems are the same as at optimality $\alpha_i = \|x^{(i)} - a^{(i)}\|_2$.

To solve problems of the form (1.2) interior point methods replace the constraint $\{\alpha, x\} \in S$ through the introduction of a *barrier function*. In particular they assume that there is a real valued function p such that as $\{\alpha, x\}$ moves towards the boundary of S the value of p goes to infinity. A popular class of interior point methods known as *path following methods* [26, 11], they consider relaxations of (1.2) of the form $\min_{\{\alpha,x\}\in\mathbb{R}^n\times\mathbb{R}^d} t \cdot 1^\top\alpha + p(\alpha, x)$. The minimizers of this function form a path, known as the central path, parameterized by t. The methods then use variants of Newton's method to follow the path until t is large enough that a high quality approximate solution is obtained. The number of iterations of these methods are then typically governed by a property of p known as its self concordance ν. Given a ν-self concordant barrier, typically interior point methods require $O(\sqrt{\nu}\log\frac{1}{\epsilon})$ iterations to compute a $(1 + \epsilon)$-approximate solution.

For our particular convex set, the construction of our barrier function is particularly simple, we consider each constraint $\|x - a^{(i)}\|_2 \leq \alpha_i$ individually. In particular, it is known that the function $p^{(i)}(\alpha, x) = -\ln\left(\alpha_i^2 - \|x - a^{(i)}\|_2^2\right)$ is a 2-self-concordant barrier function for the set $S^{(i)} = \left\{x \in \mathbb{R}^d, \alpha \in \mathbb{R}^n \,|\, \|x - a^{(i)}\|_2 \leq \alpha_i\right\}$ [21, Lem 4.3.3]. Since $\cap_{i\in[n]} S^{(i)} = S$ we can use the barrier $\sum_{i\in[n]} p^{(i)}(\alpha, x)$ for $p(\alpha, x)$ and standard self-concordance theory shows that this is an $O(n)$ self concordant barrier for S. Consequently, this easily yields an interior point method for solving the geometric median problem in $O((nd)^{O(1)}\log\frac{1}{\epsilon})$ time.

Difficulties

Unfortunately obtaining a nearly linear time algorithm for geometric median using interior point methods as presented poses numerous difficulties. Particularly troubling is the number of iterations required by standard interior point algorithms. The approach outlined in the previous section produced an $O(n)$-self concordant barrier and even if we use more advanced self concordance machinery, i.e. the universal barrier [22], the best known self concordance of barrier for the convex set $\sum_{i\in[n]} \|x - a^{(i)}\|_2 \leq c$ is $O(d)$. An interesting open question still left open by our work is to determine what is the minimal self concordance of a barrier for this set.

Consequently, even if we could implement every iteration of an interior point scheme in nearly linear time it is unclear whether one should hope for a nearly linear time interior point algorithm for the geometric median. While there are a instances of outperforming standard self-concordance analysis [20, 17], these instances are few, complex, and to varying degrees specialized to the problems they solve. Moreover, we are unaware of any interior point scheme providing a provable nearly linear time for a general nontrivial convex optimization problem.

Beyond Standard Interior Point

Despite these difficulties we do obtain a nearly linear time interior point based algorithm that only requires $O(\log \frac{n}{\epsilon})$ iterations, i.e. increases to the path parameter. After choosing the natural penalty functions $p^{(i)}$ described above, we optimize in closed form over the α_i to obtain the following penalized objective function:[3]

$$\min_x f_t(x) \quad \text{where} \quad f_t(x) = \sum_{i \in [n]} \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2} - \ln\left[1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}\right]$$

We then approximately minimize $f_t(x)$ for increasing t. We let $x_t \stackrel{\text{def}}{=} \arg\min_{x \in \mathbb{R}^d} f_t(x)$ for $x \geq 0$, and thinking of $\{x_t : t \geq 0\}$ as a continuous curve known as the *central path*, we show how to approximately follow this path. As $\lim_{t \to \infty} x_t = x_*$ this approach yields a $(1 + \epsilon)$-approximation.

So far our analysis is standard and interior point theory yields an $\Omega(\sqrt{n})$ iteration interior point scheme. To overcome this we take a more detailed look at x_t. We note that for any t if there is any rapid change in x_t it must occur in the direction of the smallest eigenvector of $\nabla^2 f_t(x)$, denoted v_t, what we henceforth may refer to as the *bad direction* at x_t. More precisely, for all directions $d \perp v_t$ it is the case that $d^\top(x_t - x_{t'})$ is small for $t' \leq ct$ for a small constant c.

In fact, we show that this movement over such a *long step,* i.e. a constant increase in t, in the directions orthogonal to the bad direction is small enough that for any movement around a ball of this size the Hessian of f_t only changes by a small multiplicative constant. In short, starting at x_t there exists a point y obtained just by moving from x_t in the bad direction, such that y is close enough to $x_{t'}$ that standard first order method will converge quickly to $x_{t'}$! Thus, we might hope to find such a y, quickly converge to $x_{t'}$ and repeat. If we increase t by a multiplicative constant in every such iterations, standard interior point theory suggests that $O(\log \frac{n}{\epsilon})$ iterations suffices.

Building an Algorithm

To turn the structural result in the previous section into a fast algorithm there are several further issues we need to address. We need to

- (1) Show how to find the point along the bad direction that is close to $x_{t'}$
- (2) Show how to solve linear systems in the Hessian to actually converge quickly to $x_{t'}$
- (3) Show how to find the bad direction
- (4) Bound the accuracy required by these computations

Deferring (1) for the moment, our solution to the rest are relatively straightforward. Careful inspection of the Hessian of f_t reveals that it is well approximated by a multiple of the identity matrix minus a rank 1 matrix. Consequently using explicit formulas for the inverse of of matrix under rank 1 updates, i.e. the Sherman-Morrison formula, we can solve such systems in nearly linear time thereby addressing (2). For (3), we show that the well known power method carefully applied to the Hessian yields the bad direction if it exists. Finally, for (4) we show that a constant approximate geometric median is near enough to the central path for $t = \Theta(\frac{1}{f(x_*)})$ and that it suffices to compute a central path point at $t = O(\frac{n}{f(x_*)\epsilon})$ to compute a $1 + \epsilon$-geometric median. Moreover, for these values of t, the precision needed in other operations is clear.

[3] It is unclear how to extend our proof for the simpler function: $\sum_{i \in [n]} \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}$.

The more difficult operation is (1). Given x_t and the bad direction exactly, it is still not clear how to find the point along the bad direction line from x_t that is close to $x_{t'}$. Just performing binary search on the objective function a priori might not yield such a point due to discrepancies between a ball in Euclidean norm and a ball in hessian norm and the size of the distance from the optimal point in euclidean norm. To overcome this issue we still line search on the bad direction, however rather than simply using $f(x_t + \alpha \cdot v_t)$ as the objective function to line search on, we use the function $g(\alpha) = \min_{\|x - x_t - \alpha \cdot v_t\|_2 \leq c} f(x)$ for some constant c, that is given an α we move α in the bad direction and take the best objective function value in a ball around that point. For appropriate choice of c the minimizers of α will include the optimal point we are looking for. Moreover, we can show that g is convex and that it suffices to perform the minimization approximately.

Putting these pieces together yields our result. We perform $O(\log \frac{n}{\epsilon})$ iterations of interior point (i.e. increasing t), where in each iteration we spend $O(nd \log \frac{n}{\epsilon})$ time to compute a high quality approximation to the bad direction, and then we perform $O(\log \frac{n}{\epsilon})$ approximate evaluations on $g(\alpha)$ to binary search on the bad direction line, and then to approximately evaluate g we perform gradient descent in approximate Hessian norm to high precision which again takes $O(nd \log \frac{n}{\epsilon})$ time. Altogether this yields a $O(nd \log^3 \frac{n}{\epsilon})$ time algorithm to compute a $1 + \epsilon$ geometric median. Here we made minimal effort to improve the log factors and plan to investigate this further in future work.

1.3 Overview of $O(d\epsilon^{-2})$ Time Algorithm

In addition to providing a nearly linear time algorithm we provide a stand alone result on quickly computing a crude $(1+\epsilon)$-approximate geometric median in Section C. In particular, given an oracle for sampling a random $a^{(i)}$ we provide an $O(d\epsilon^{-2})$, i.e. sublinear, time algorithm that computes such an approximate median. Our algorithm for this result is fairly straightforward. First, we show that random sampling can be used to obtain some constant approximate information about the optimal point in constant time. In particular we show how this can be used to deduce an Euclidean ball which contains the optimal point. Second, we perform stochastic subgradient descent within this ball to achieve our desired result.

1.4 Paper Organization

The rest of the paper is structured as follows. After covering preliminaries in Section 2, in Section 3 we provide various results about the central path that we use to derive our nearly linear time algorithm. In Section 4 we then provide our nearly linear time algorithm. All the proofs and supporting lemmas for these sections are deferred to Appendix A and Appendix B. In Appendix C we provide our $O(d/\epsilon^2)$ algorithm, in Appendix D we provide the derivation of our penalized objective function, in Appendix E we provide general technical machinery we use throughout and in Appendix F we show how to extend our results to Weber's problem, i.e. weighted geometric median.

2 Notation

2.1 General Notation

We use bold to denote a matrix. For a symmetric positive semidefinite matrix (PSD), $\mathbf{A}$, we let $\lambda_1(\mathbf{A}) \geq ... \geq \lambda_n(\mathbf{A}) \geq 0$ denote the eigenvalues of $\mathbf{A}$ and let $v_1(\mathbf{A}), ..., v_n(\mathbf{A})$ denote corresponding eigenvectors. We let $\|x\|_{\mathbf{A}} \stackrel{\text{def}}{=} \sqrt{x^\top \mathbf{A} x}$ and for PSD we use $\mathbf{A} \preceq \mathbf{B}$ and $\mathbf{B} \preceq \mathbf{A}$ to denote the conditions that $x^\top \mathbf{A} x \leq x^\top \mathbf{B} x$ for all x and $x^\top \mathbf{B} x \leq x^\top \mathbf{A} x$ for all x respectively.

2.2 Problem Notation

The central problem of this paper is as follows: we are given points $a^{(1)}, ..., a^{(n)} \in \mathbb{R}^d$ and we wish to compute a geometric median, i.e. $x_* \in \arg\min_{x \in \mathbb{R}^d} f(x)$ where $f(x) = \sum_{i \in [n]} \|a^{(i)} - x\|_2$. We call a point $x \in \mathbb{R}^d$ an $(1+\epsilon)$-approximate geometric median if $f(x) \leq (1+\epsilon) f(x_*)$.

2.3 Penalized Objective Notation

To solve this problem, we smooth the objective function f and instead consider the following family of *penalized objective functions* parameterized by $t > 0$

$$\min_{x \in \mathbb{R}^d} f_t(x) \quad \text{where} \quad f_t(x) = \sum_{i \in [n]} \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2} - \ln\left[1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}\right]$$

This penalized objective function is derived from a natural interior point formulation of the geometric median problem (See Section D). For all *path parameters* $t > 0$, we let $x_t \stackrel{\text{def}}{=} \arg\min_x f_t(x)$. Our primary goal is to obtain good approximations to the *central path* $\{x_t : t > 0\}$ for increasing values of t.

We let $g_t^{(i)}(x) \stackrel{\text{def}}{=} \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}$ and $f_t^{(i)}(x) \stackrel{\text{def}}{=} g_t^{(i)}(x) - \ln(1 + g_t^{(i)}(x))$ so $f_t(x) = \sum_{i \in [n]} f_t^{(i)}(x)$. We refer to the quantity $w_t(x) \stackrel{\text{def}}{=} \sum_{i \in [n]} \frac{1}{1 + g_t^{(i)}(x)}$ as *weight* as it is a natural measure of total contribution of the $a^{(i)}$ to $\nabla^2 f_t(x)$. We let

$$\bar{g}_t(x) \stackrel{\text{def}}{=} w_t(x) \left[\sum_{i \in [n]} \frac{1}{(1 + g_t^{(i)}(x_t)) g_t^{(i)}(x_t)}\right]^{-1} = \frac{\sum_{i \in [n]} \frac{1}{1 + g_t^{(i)}(x_t)}}{\sum_{i \in [n]} \frac{1}{(1 + g_t^{(i)}(x_t)) g_t^{(i)}(x_t)}}$$

denote a weighted harmonic mean of g that helps upper bound the rate of change of the central path. Furthermore, we let $u^{(i)}(x)$ denote the unit vector corresonding to $x - a^{(i)}$, i.e. $u^{(i)}(x) \stackrel{\text{def}}{=} x - a^{(i)} / \|x - a^{(i)}\|_2$ when $\|x - a^{(i)}\|_2 \neq 0$ and $u^{(i)}(x) \stackrel{\text{def}}{=} 0$ otherwise. Finally we let $\mu_t(x) \stackrel{\text{def}}{=} \lambda_d(\nabla^2 f_t(x))$ denote the minimum eigenvalue of $\nabla^2 f_t(x)$, and let $v_t(x)$ denote a corresponding eigenvector. To simplify notation we often drop the (x) in these definitions when $x = x_t$ and t is clear from context.

3 Properties of the Central Path

Here provide various facts regarding the penalized objective function and the central path. While we use the lemmas in this section throughout the paper, the main contribution of this section is Lemma 5 in Section 3.3. There we prove that with the exception of a single direction, the change in the central path is small over a constant multiplicative change in the path parameter. In addition, we show that our penalized objective function is stable under changes in a $O(\frac{1}{t})$ Euclidean ball (Section 3.1), we bound the change in the Hessian over the central path (Section 3.2), and we relate $f(x_t)$ to $f(x_*)$ (Section 3.4).

3.1 How Much Does the Hessian Change in General?

Here, we show that the Hessian of the penalized objective function is stable under changes in a $O(\frac{1}{t})$ sized Euclidean ball. This shows that if we have a point which is close to a central path point in Euclidean norm, then we can use Newton method to find it.

Lemma 1. *Suppose that* $\|x-y\|_2 \leq \frac{\epsilon}{t}$ *with* $\epsilon \leq \frac{1}{20}$. *Then, we have*

$$(1-6\epsilon^{2/3})\nabla^2 f_t(x) \preceq \nabla^2 f_t(y) \preceq (1+6\epsilon^{2/3})\nabla^2 f_t(x).$$

3.2 How Much Does the Hessian Change Along the Path?

Here we bound how much the Hessian of the penalized objective function can change along the central path. First we provide the following lemma bound several aspects of the penalized objective function and proving that the weight, w_t, only changes by a small amount multiplicatively given small multiplicative changes in the path parameter, t.

Lemma 2. *For all* $t \geq 0$ *and* $i \in [n]$ *the following hold*

$$\left\|\frac{d}{dt}x_t\right\|_2 \leq \frac{1}{t^2}\bar{g}_t(x_t) \quad , \quad \left|\frac{d}{dt}g_t^{(i)}(x_t)\right| \leq \frac{1}{t}\left(g_t^{(i)}(x_t)+\bar{g}_t\right) \quad , \text{ and } \quad \left|\frac{d}{dt}w_t\right| \leq \frac{2}{t}w_t$$

Consequently, for all $t' \geq t$ *we have that* $\left(\frac{t}{t'}\right)^2 w_t \leq w_{t'} \leq \left(\frac{t'}{t}\right)^2 w_t$.

Next we use this lemma to bound the change in the Hessian with respect to t.

Lemma 3. *For all* $t \geq 0$ *we have*

$$-12\cdot t\cdot w_t\mathbf{I} \preceq \frac{d}{dt}\left[\nabla^2 f_t(x_t)\right] \preceq 12\cdot t\cdot w_t\mathbf{I} \tag{3.1}$$

and therefore for all $\beta \in [0, \frac{1}{8}]$

$$\nabla^2 f(x_t) - 15\beta t^2 w_t\mathbf{I} \preceq \nabla^2 f(x_{t(1+\beta)}) \preceq \nabla^2 f(x_t) + 15\beta t^2 w_t\mathbf{I}\,. \tag{3.2}$$

3.3 Where is the Next Optimal Point?

Here we prove our main result of this section. We prove that over a long step the central path moves very little in directions orthogonal to the smallest eigenvector of the Hessian. We begin by noting the Hessian is approximately a scaled identity minus a rank 1 matrix.

Lemma 4. *For all* t, *we have*

$$\frac{1}{2}\left[t^2\cdot w_t\mathbf{I} - (t^2\cdot w_t - \mu_t)v_t v_t^\top\right] \preceq \nabla^2 f_t(x_t) \preceq t^2\cdot w_t\mathbf{I} - (t^2\cdot w_t-\mu_t)v_t v_t^\top.$$

Using this and the lemmas of the previous section we bound the amount x_t can move in every direction far from v_t.

Lemma 5 (The Central Path is Almost Straight)**.** *For all* $t \geq 0$, $\beta \in [0, \frac{1}{600}]$, *and any unit vector* y *with* $|\langle y, v_t\rangle| \leq \frac{1}{t^2\cdot\kappa}$ *where* $\kappa = \max_{\delta\in[t,(1+\beta)t]}\frac{w_\delta}{\mu_\delta}$, *we have* $y^\top(x_{(1+\beta)t} - x_t) \leq \frac{6\beta}{t}$.

3.4 Where is the End?

In this section, we bound the quality of the central path with respect to the geometric median objective. In particular, we show that if we can solve the problem for some $t = \frac{2n}{\epsilon f(x_*)}$ then we obtain an $(1+\epsilon)$-approximate solution. As our algorithm ultimately starts from an initial $t = 1/O(f(x_*))$ and increases t by a multiplicative constant in every iteration, this yields an $O(\log\frac{n}{\epsilon})$ iteration algorithm.

Lemma 6. $f(x_t) - f(x_*) \leq \frac{2n}{t}$ *for all* $t > 0$.

4 Nearly Linear Time Geometric Median

Here we show how to use the structural results from the previous section to obtain a nearly linear time algorithm for computing the geometric median. Our algorithm follows a simple structure (See Algorithm 1). First we use simply average the $a^{(i)}$ to compute a 2-approximate median, denoted $x^{(0)}$. Then for a number of iterations we repeatedly move closer to x_t for some path parameter t, compute the minimum eigenvector of the Hessian, and line search in that direction to find an approximation to a point further along the central path. Ultimately, this yields a point $x^{(k)}$ that is precise enough approximation to a point along the central path with large enough t that we can simply out $x^{(k)}$ as our $(1+\epsilon)$-approximate geometric median.

Algorithm 1: `AccurateMedian`(ϵ)

```
**Input**: points $a^{(1)}, ..., a^{(n)} \in \mathbb{R}^d$
**Input**: desired accuracy $\epsilon \in (0,1)$

// Compute a 2-approximate geometric median and use it to center
Compute $x^{(0)} := \frac{1}{n}\sum_{i\in[n]} a^{(i)}$ and $\widetilde{f}_* := f(x^{(0)})$ // Note $\widetilde{f}_* \leq 2f(x_*)$ by Lemma 18
Let $t_i = \frac{1}{400\widetilde{f}_*}(1+\frac{1}{600})^{i-1}$, $\tilde{\epsilon}_* = \frac{1}{3}\epsilon$, and $\tilde{t}_* = \frac{2n}{\tilde{\epsilon}_*\cdot\tilde{f}_*}$ .
Let $\epsilon_v = \frac{1}{8}(\frac{\tilde{\epsilon}_*}{7n})^2$ and let $\epsilon_c = (\frac{\epsilon_v}{36})^{\frac{3}{2}}$ .
$x^{(1)} =$ LineSearch$(x^{(0)}, t_1, t_1, 0, \epsilon_c)$ .

// Iteratively improve quality of approximation
Let $k = \max_{i\in\mathbb{Z}} t_i \leq \tilde{t}_*$
**for** $i \in [1,k]$ **do**
    // Compute $\epsilon_v$-approximate minimum eigenvalue and eigenvector of $\nabla^2 f_{t_i}(x^{(i)})$
    $(\lambda^{(i)}, u^{(i)}) =$ ApproxMinEig$(x^{(i)}, t_i, \epsilon_v)$ .

    // Line search to find $x^{(i+1)}$ such that $\|x^{(i+1)} - x_{t_{i+1}}\|_2 \leq \frac{\epsilon_c}{t_{i+1}}$
    $x^{(i+1)} =$ LineSearch$(x^{(i)}, t_i, t_{i+1}, u^{(i)}, \epsilon_c)$ .
**end**
**Output:** $\epsilon$-approximate geometric median $x^{(k+1)}$.
```

We split the remainder of the algorithm specification and its analysis into several parts. First in Section 4.1 we show how to compute an approximate minimum eigenvector and eigenvalue of the Hessian of the penalized objective function. Then in Section 4.2 we show how to use this eigenvector to line search for the next central path point. Finally, in Section 4.3 we put these results together to obtain our nearly linear time algorithm. Throughout this section we will want an upper bound to $f(x_*)$ and a slight lower bound on ϵ, the geometric median accuracy we are aiming for. We use an easily computed $\tilde{f}_* \leq 2f(x_*)$ for the former and $\tilde{\epsilon}_* = \frac{1}{3}\epsilon$ throughout the section.

4.1 Eigenvector Computation and Hessian Approximation

Here we show how to compute the minimum eigenvector of $\nabla^2 f_t(x)$ and thereby obtain a concise approximation to $\nabla^2 f_t(x)$. Our main algorithmic tool is the well known power method and the fact that it converges quickly on a matrix with a large eigenvalue gap. To improve our logarithmic terms we need a slightly non-standard analysis of the method and therefore we provide and analyze this method for completeness in Section B.1. Using this tool we estimate the top eigenvector as follows.

```
Algorithm 2: ApproxMinEig($x, t, \epsilon$)
    Input: Point $x \in \mathbb{R}^d$, path parameter $t$, and target accuracy $\epsilon$.
    Let $\mathbf{A} = \sum_{i\in[n]} \frac{t^4(x-a^{(i)})(x-a^{(i)})^\top}{(1+g_t^{(i)}(x))^2 g_t^{(i)}(x)}$
    Let $u :=$ PowerMethod($\mathbf{A}, \Theta(\log\left(\frac{n}{\epsilon}\right))$)
    Let $\lambda = u^\top \nabla^2 f_t(x) u$
    Output: $(\lambda, u)$
```

Lemma 7 (Computing Hessian Approximation). *Let $x \in \mathbb{R}^d$, $t > 0$, and $\epsilon \in (0, \frac{1}{4})$. The algorithm* `ApproxMinEig`$(x, t, \epsilon)$ *outputs (λ, u) in $O(nd\log\frac{n}{\epsilon})$ time such that if $\mu_t(x) \leq \frac{1}{4}t^2 w_t(x)$ then $\langle v_t(x), u\rangle^2 \geq 1 - \epsilon$ with high probability in n/ϵ. Furthermore, if $\epsilon \leq \left(\frac{\mu_t(x)}{8t^2 \cdot w_t(x)}\right)^2$ then $\frac{1}{4}\mathbf{Q} \preceq \nabla^2 f_t(x) \preceq 4\mathbf{Q}$ with high probability in n/ϵ where $\mathbf{Q} \overset{\text{def}}{=} t^2 \cdot w_t(x) - \left(t^2 \cdot w_t(x) - \lambda\right) uu^\top$.*

Furthermore, we show that the $v^{(i)}$ computed by this algorithm is sufficiently close to the bad direction. Combining 7 with the structural results from the previous section and Lemma 29, a minor technical lemma regarding the transitivity of large inner products,we provide the following lemma.

Lemma 8. *Let $(\lambda, u) =$* `ApproxMinEig`(x, t, ϵ_v) *for $\epsilon_v < \frac{1}{8}$ and $\|x - x_t\|_2 \leq \frac{\epsilon_c}{t}$ for $\epsilon_c \leq (\frac{\epsilon_v}{36})^{\frac{3}{2}}$. If $\mu_t \leq \frac{1}{4}t^2 \cdot w_t$ then with high probability in n/ϵ_v for all unit vectors $y \perp u$, we have $\langle y, v_t\rangle^2 \leq 8\epsilon_v$.*

Note that this lemma assumes μ_t is small. When μ_t is large, we instead show that the next central path point is close to the current point and hence we do not need to compute the bad direction to center quickly.

Lemma 9. *Suppose $\mu_t \geq \frac{1}{4}t^2 \cdot w_t$ and let $t' \in [t, (1 + \frac{1}{600})t]$ then $\|x_{t'} - x_t\|_2 \leq \frac{1}{100t}$.*

4.2 Line Searching

Here we show how to line search along the bad direction to find the next point on the central path. Unfortunately, simply performing binary search on objective function directly may not suffice. If we search over α to minimize $f_{t_{i+1}}(y^{(i)} + \alpha v^{(i)})$ it is unclear if we actually obtain a point close to x_{t+1}. It might be the case that even after minimizing α we would be unable to move towards x_{t+1} efficiently.

To overcome this difficulty, we use the fact that over the region $\|x - y\|_2 = O(\frac{1}{t})$ the Hessian changes by at most a constant and therefore we can minimize $f_t(x)$ over this region extremely quickly. Therefore, we instead line search on the following function

$$g_{t,y,v}(\alpha) \overset{\text{def}}{=} \min_{\|x-(y+\alpha v)\|_2 \leq \frac{1}{49t}} f_t(x) \tag{4.1}$$

and use that we can evaluate $g_{t,y,v}(\alpha)$ approximately by using an appropriate centering procedure. We can show (See Lemma 31) that $g_{t,y,v}(\alpha)$ is convex and therefore we can minimize it efficiently just by doing an appropriate binary search. By finding the approximately minimizing α and outputting the corresponding approximately minimizing x, we can obtain $x^{(i+1)}$ that is close enough to $x_{t_{i+1}}$. For notational convenience, we simply write $g(\alpha)$ if t, y, v is clear from the context.

First, we show how we can locally center and provide error analysis for that algorithm.

```
Algorithm 3: LocalCenter$(y, t, \epsilon)$
    Input: Point $y \in \mathbb{R}^d$, path parameter $t > 0$, target accuracy $\epsilon > 0$.
    Let $(\lambda, v) :=$ ApproxMinEig$(x, t, \epsilon)$.
    Let $\mathbf{Q} = t^2 \cdot w_t(y)\mathbf{I} - \left(t^2 \cdot w_t(y) - \lambda\right) vv^\top$
    Let $x^{(0)} = y$
    for $i = 1, ..., k = 64 \log \frac{1}{\epsilon}$ do
        Let $x^{(i)} = \min_{\|x-y\|_2 \leq \frac{1}{49t}} f(x^{(i-1)}) + \langle \nabla f_t(x^{(i-1)}), x - x^{(i-1)} \rangle + 4\|x - x^{(i-1)}\|_{\mathbf{Q}}^2$.
    end
    Output: $x^{(k)}$
```

Lemma 10. *Given some $y \in \mathbb{R}^d$, $t > 0$ and $0 \leq \epsilon \leq \left(\frac{\mu_t(x)}{8t^2 \cdot w_t(x)}\right)^2$. In $O(nd \log(\frac{n}{\epsilon}))$ time* `LocalCenter`(y, t, ϵ) *computes $x^{(k)}$ such that with high probability in n/ϵ.*

$$f_t(x^{(k)}) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x) \leq \epsilon \left(f_t(y) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x) \right) .$$

Using this local centering algorithm as well as a general result for minimizing one dimensional convex functions using a noisy oracle (See Section E.3) we obtain our line search algorithm.

```
Algorithm 4: LineSearch$(y, t, t', u, \epsilon)$
    Input: Point $y \in \mathbb{R}^d$, current path parameter $t$, next path parameter $t'$, bad direction $u$,
    target accuracy $\epsilon$
    Let $\epsilon_O = \left(\frac{\epsilon \tilde{\epsilon}_*}{160n^2}\right)^2$, $\ell = -6\widetilde{f}_*$, $u = 6\widetilde{f}_*$.
    Define the oracle $q : \mathbb{R} \to \mathbb{R}$ by $q(\alpha) = f_{t'}$ (LocalCenter $(y + \alpha u, t', \epsilon_O)$)
    Let $\alpha' =$ OneDimMinimizer$(\ell, u, \epsilon_O, q, t'n)$
    Output: $x' =$ LocalCenter $(y + \alpha u, t', \epsilon_O)$
```

Lemma 11. *Let $\frac{1}{400f(x_*)} \leq t \leq t' \leq (1 + \frac{1}{600})t \leq \frac{2n}{\tilde{\epsilon}_* \cdot \tilde{f}_*}$ and let $(\lambda, u) =$* `ApproxMinEig`(y, t, ϵ_v) *for $\epsilon_v \leq \frac{1}{8}(\frac{\tilde{\epsilon}_*}{3n})^2$ and $y \in \mathbb{R}^d$ such that $\|y - x_t\|_2 \leq \frac{1}{t}(\frac{\epsilon_v}{36})^{\frac{3}{2}}$. In $O(nd \log^2(\frac{n}{\tilde{\epsilon}_* \cdot \epsilon \cdot \epsilon_v}))$ time and $O(\log(\frac{n}{\tilde{\epsilon}_* \cdot \epsilon}))$ calls to the* `LocalCenter`, `LineSearch`(y, t, t', u, ϵ) *outputs x' such that $\|x' - x_{t'}\|_2 \leq \frac{\epsilon}{t'}$ with high probability in n/ϵ.*

We also provide the following lemma useful for finding the first center.

Lemma 12. *Let $\frac{1}{400f(x_*)} \leq t \leq t' \leq (1 + \frac{1}{600})t \leq \frac{2n}{\tilde{\epsilon}_* \cdot \tilde{f}_*}$ and let $x \in \mathbb{R}^d$ satisfy $\|x - x_t\|_2 \leq \frac{1}{100t}$. Then, in $O(nd \log^2(\frac{n}{\epsilon \cdot \tilde{\epsilon}_*}))$ time,* `LineSearch`(x, t, t, u, ϵ) *outputs y such that $\|y - x_t\|_2 \leq \frac{\epsilon}{t}$ for any vector $u \in \mathbb{R}^d$.*

4.3 Putting It All Together

Combining the results of the previous sections, we prove our main theorem.

Theorem 1. *In $O(nd \log^3(\frac{n}{\epsilon}))$ time, Algorithm 1 outputs an $(1 + \epsilon)$-approximate geometric median with constant probability.*

5 Acknowledgments

We thank Yan Kit Chim, Ravi Kannan, and Jonathan A. Kelner for many helpful conversations. We thank the reviewers for their help in completing the previous work table. This work was partially supported by NSF awards 0843915, 1065106 and 1111109, NSF Graduate Research Fellowship (grant no. 1122374) and Sansom Graduate Fellowship in Computer Science. Part of this work was done while authors were visiting the Simons Institute for the Theory of Computing, UC Berkeley.

References

[1] Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 250–257, 2002.

[2] Chanderjit Bajaj. The algebraic degree of geometric optimization problems. *Discrete & Computational Geometry*, 3(2):177–191, 1988.

[3] Egon Balas and Chang-Sung Yu. A note on the weiszfeld-kuhn algorithm for the general fermat problem. *Managme Sci Res Report*, (484):1–6, 1982.

[4] Prosenjit Bose, Anil Maheshwari, and Pat Morin. Fast approximations for sums of distances, clustering and the Fermat-Weber problem. *Computational Geometry*, 24(3):135 – 146, 2003.

[5] Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

[6] R. Chandrasekaran and A. Tamir. Open questions concerning weiszfeld's algorithm for the fermat-weber location problem. *Mathematical Programming*, 44(1-3):293–295, 1989.

[7] Hui Han Chin, Aleksander Madry, Gary L. Miller, and Richard Peng. Runtime guarantees for regression problems. In *ITCS*, pages 269–282, 2013.

[8] Leon Cooper and I.Norman Katz. The weber problem revisited. *Computers and Mathematics with Applications*, 7(3):225 – 234, 1981.

[9] Zvi Drezner, Kathrin Klamroth, Anita SchÃPbel, and George Wesolowsky. *Facility location*, chapter The Weber problem, pages 1–36. Springer, 2002.

[10] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.

[11] Clovis C Gonzaga. Path-following methods for linear programming. *SIAM review*, 34(2):167–224, 1992.

[12] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 126–134. ACM, 2005.

[13] P. Indyk and Stanford University. Computer Science Dept. *High-dimensional computational geometry*. Stanford University, 2000.

[14] Jakob Krarup and Steven Vajda. On torricelli's geometrical solution to a problem of fermat. *IMA Journal of Management Mathematics*, 8(3):215–224, 1997.

[15] Richard A. Kronmal and Arthur V. Peterson. The alias and alias-rejection-mixture methods for generating random variables from probability distributions. In *Proceedings of the 11th Conference on Winter Simulation - Volume 1*, WSC '79, pages 269–280, Piscataway, NJ, USA, 1979. IEEE Press.

[16] HaroldW. Kuhn. A note on fermat's problem. *Mathematical Programming*, 4(1):98–107, 1973.

[17] Yin Tat Lee and Aaron Sidford. Path-finding methods for linear programming : Solving linear programs in õ(sqrt(rank)) iterations and faster algorithms for maximum flow. In *55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014, 18-21 October, 2014, Philadelphia, PA, USA*, pages 424–433, 2014.

[18] Hendrik P. Lopuhaa and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1):229–248, 03 1991.

[19] Hendrik P Lopuhaa and Peter J Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, pages 229–248, 1991.

[20] Aleksander Madry. Navigating central path with electrical flows: from flows to matchings, and back. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, 2013.

[21] Yu Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume I. 2003.

[22] Yurii Nesterov and Arkadii Semenovich Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Society for Industrial and Applied Mathematics, 1994.

[23] Lawrence M. Ostresh. On the convergence of a class of iterative methods for solving the weber location problem. *Operations Research*, 26(4):597–609, 1978.

[24] Pablo A. Parrilo and Bernd Sturmfels. Minimizing polynomial functions. In *DIMACS Workshop on Algorithmic and Quantitative Aspects of Real Algebraic Geometry in Mathematics and Computer Science, March 12-16, 2001, DIMACS Center, Rutgers University, Piscataway, NJ, USA*, pages 83–100, 2001.

[25] Frank Plastria and Mohamed Elosmani. On the convergence of the weiszfeld algorithm for continuous single facility location allocation problems. *TOP*, 16(2):388–406, 2008.

[26] James Renegar. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical Programming*, 40(1-3):59–93, 1988.

[27] Yehuda Vardi and Cun-Hui Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

[28] Vincenzo Viviani. De maximis et minimis geometrica divinatio liber 2. *De Maximis et Minimis Geometrica Divinatio*, 1659.

[29] Alfred Weber. *The Theory of the Location of Industries.* Chicago University Press, 1909. Aber den I der Industrien.

[30] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Mathematical Journal*, pages 355–386, 1937.

[31] Guoliang Xue and Yinyu Ye. An efficient algorithm for minimizing a sum of euclidean norms with applications. *SIAM Journal on Optimization*, 7:1017–1036, 1997.

[32] Yinyu Ye. *Interior point algorithms: theory and analysis*, volume 44. John Wiley & Sons, 2011.

A Properties of the Central Path (Proofs)

Here we provide proofs of the claims in Section 3 as well as additional technical lemmas we use throughout the paper.

A.1 Basic Facts

Here we provide basic facts regarding the central path that we will use throughout our analysis. First we compute various derivatives of the penalized objective function.

Lemma 13 (Path Derivatives)**.** *We have*

$$\nabla f_t(x) = \sum_{i \in [n]} \frac{t^2 (x - a^{(i)})}{1 + g_t^{(i)}(x)} \quad , \quad \nabla^2 f_t(x) = \sum_{i \in [n]} \frac{t^2}{1 + g_t^{(i)}(x)} \left(\mathbf{I} - \frac{t^2 (x - a^{(i)})(x - a^{(i)})^\top}{g_t^{(i)}(x)(1 + g_t^{(i)}(x))} \right) \quad \text{, and}$$

$$\frac{d}{dt} x_t = -\left(\nabla^2 f_t(x_t) \right)^{-1} \sum_{i \in [n]} \frac{t(x_t - a^{(i)})}{(1 + g_t^{(i)}(x_t)) g_t^{(i)}(x_t)}$$

Proof of Lemma 13. Direct calculation shows that

$$\begin{aligned} \nabla f_t^{(i)}(x) &= \frac{t^2 (x - a^{(i)})}{\sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} - \frac{1}{1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} \left(\frac{t^2 (x - a^{(i)})}{\sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} \right) \\ &= \frac{t^2 (x - a^{(i)})}{1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} = \frac{t^2 (x - a^{(i)})}{1 + g_t^{(i)}(x)} \end{aligned}$$

and

$$\begin{aligned} \nabla^2 f_t^{(i)}(x) &= \frac{t^2}{1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} \mathbf{I} - \left(\frac{1}{1 + \sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} \right)^2 \frac{t^4 (x - a^{(i)})(x - a^{(i)})^\top}{\sqrt{1 + t^2 \|x - a^{(i)}\|_2^2}} \\ &= \frac{t^2}{1 + g_t^{(i)}(x)} \left(\mathbf{I} - \frac{t^2 (x - a^{(i)})(x - a^{(i)})^\top}{g_t^{(i)}(x)(1 + g_t^{(i)}(x))} \right) \end{aligned}$$

and

$$\begin{aligned}\left(\frac{d}{dt}\nabla f_t^{(i)}\right)(x) &= \frac{2t(x-a^{(i)})}{1+\sqrt{1+t^2\|x-a^{(i)}\|_2^2}} - \frac{t^2\cdot(x-a^{(i)})\cdot t\|x-a^{(i)}\|_2^2}{\left(1+\sqrt{1+t^2\|x-a^{(i)}\|}\right)^2\sqrt{1+t^2\|x-a^{(i)}\|_2^2}} \\ &= \frac{t\cdot(x-a^{(i)})}{1+g_t^{(i)}(x)}\left(2-\frac{g_t^{(i)}(x)^2-1}{(1+g_t^{(i)}(x))g_t^{(i)}(x)}\right) \\ &= \frac{t\cdot(x-a^{(i)})}{1+g_t^{(i)}(x)}\left(\frac{2g_t^{(i)}(x)-(g_t^{(i)}(x)-1)}{g_t^{(i)}(x)}\right) = \frac{t\cdot(x-a^{(i)})}{g_t^{(i)}(x)}\end{aligned}$$

Finally, by the optimality of x_t we have that $\nabla f_t(x_t) = 0$. Consequently,

$$\nabla^2 f_t(x_t)\frac{d}{dt}x_t + \left(\frac{d}{dt}\nabla f_t\right)(x_t) = 0.$$

and solving for $\frac{d}{dt}x_t$ then yields

$$\begin{aligned}\frac{d}{dt}x_t &= -\left(\nabla^2 f_t(x_t)\right)^{-1}\left(\left(\frac{d}{dt}\nabla f_t\right)(x_t)\right) \\ &= -\left(\nabla^2 f_t(x_t)\right)^{-1}\left(\left(\frac{d}{dt}\nabla f_t\right)(x_t) - \frac{1}{t}\nabla f_t(x_t)\right) \\ &= -\left(\nabla^2 f_t(x_t)\right)^{-1}\left(\sum_{i\in[n]}\left[\frac{t}{g_t^{(i)}} - \frac{t}{1+g_t^{(i)}}\right](x_t-a^{(i)})\right).\end{aligned}$$

□

Next, in we provide simple facts regarding the Hessian of the penalized objective function.

Lemma 14. *For all $t > 0$ and $x \in \mathbb{R}^d$*

$$\nabla^2 f_t(x) = \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}(x)}\left(\mathbf{I} - \left(1-\frac{1}{g_t^{(i)}(x)}\right)u^{(i)}(x)u^{(i)}(x)^\top\right)$$

and therefore

$$\sum_{i\in[n]}\frac{t^2}{(1+g_t^{(i)}(x))g_t^{(i)}(x)}\mathbf{I} \preceq \nabla^2 f_t(x) \preceq \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}(x)}\mathbf{I}$$

Proof of Lemma 14. We have that

$$\begin{aligned}\nabla^2 f_t(x) &= \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}(x)}\left(\mathbf{I} - \frac{t^2(x-a^{(i)})(x-a^{(i)})^\top}{g_t^{(i)}(x)(1+g_t^{(i)}(x))}\right) \\ &= \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}(x)}\left(\mathbf{I} - \frac{t^2\|x-a^{(i)}\|_2^2}{(1+g_t^{(i)}(x))g_t^{(i)}(x)}u^{(i)}(x)u^{(i)}(x)^\top\right).\end{aligned}$$

Since

$$\frac{t^2\|x-a^{(i)}\|_2^2}{(1+g_t^{(i)}(x))g_t^{(i)}(x)} = \frac{g_t^{(i)}(x)^2-1}{g_t^{(i)}(x)(1+g_t^{(i)}(x))} = 1-\frac{1}{g_t^{(i)}(x)}$$

the result follows. □

A.2 Stability of Hessian

Here we show that moving a point $x \in \mathbb{R}^d$ in ℓ_2, does not change the Hessian, $\nabla^2 f_t(x)$, too much spectrally. First we show that such changes do not change $g_t^{(i)}(x)$ by too much (Lemma 15) and then we use this to prove the claim, i.e. we prove Lemma 1.

Lemma 15 (Stability of g)**.** *For all $x, y \in \mathbb{R}^d$ and $t > 0$, we have*

$$g_t^{(i)}(x) - t\|x-y\|_2 \leq g_t^{(i)}(y) \leq g_t^{(i)}(x) + t\|x-y\|_2$$

Proof of Lemma 15. Direct calculation reveals that

$$\begin{aligned} g_t^{(i)}(y)^2 &= 1 + t^2\|x - a^{(i)} + y - x\|_2^2 \\ &= 1 + t^2\|x - a^{(i)}\|_2^2 + 2t^2(x - a^{(i)})^\top(y-x) + t^2\|y-x\|_2^2 \\ &= g_t^{(i)}(x)^2 + 2t^2(x - a^{(i)})^\top(y-x) + t^2\|y-x\|_2^2 \,. \end{aligned}$$

Consequently by Cauchy Schwarz

$$\begin{aligned} g_t^{(i)}(y)^2 &\leq g_t^{(i)}(x)^2 + 2t^2\|x - a^{(i)}\|_2 \cdot \|y-x\|_2 + t^2\|y-x\|_2^2 \\ &\leq \left(g_t^{(i)}(x) + t\|y-x\|_2\right)^2 \end{aligned}$$

and

$$\begin{aligned} g_t^{(i)}(y)^2 &\geq g_t^{(i)}(x)^2 - 2t^2\|x - a^{(i)}\|_2 \cdot \|y-x\|_2 + t^2\|y-x\|_2^2 \\ &\geq \left(g_t^{(i)}(x) - t\|y-x\|_2\right)^2 . \end{aligned}$$

□

Lemma 1. *Suppose that $\|x-y\|_2 \leq \frac{\epsilon}{t}$ with $\epsilon \leq \frac{1}{20}$. Then, we have*

$$(1 - 6\epsilon^{2/3})\nabla^2 f_t(x) \preceq \nabla^2 f_t(y) \preceq (1 + 6\epsilon^{2/3})\nabla^2 f_t(x).$$

Proof of Lemma 1. Here we prove the following stronger statement, for all $i \in [n]$

$$(1 - 6\epsilon^{2/3})\nabla^2 f_t^{(i)}(x) \preceq \nabla^2 f_t^{(i)}(y) \preceq (1 + 6\epsilon^{2/3})\nabla^2 f_t^{(i)}(x) \,.$$

Without loss of generality let $y - x = \alpha v + \beta u^{(i)}(x)$ for some $v \perp u^{(i)}(x)$ with $\|v\|_2 = 1$. Since $\|x-y\|_2^2 \leq \frac{\epsilon^2}{t^2}$, we know that $\alpha^2, \beta^2 \leq \frac{\epsilon^2}{t^2}$. Also, let $\bar{x} = x + \beta u^{(i)}(x)$, so that clearly, $u^{(i)}(x) = u^{(i)}(\bar{x})$. Now some manipulation reveals that for all unit vectors $z \in \mathbb{R}^d$ the following holds (so long as $u^{(i)}(x) \neq 0$ and $u^{(i)}(y) \neq 0$)

$$\begin{aligned} &\left|\left[u^{(i)}(x)^\top z\right]^2 - \left[u^{(i)}(y)^\top z\right]^2\right| \\ &= \left|\left[u^{(i)}(\bar{x})^\top z\right]^2 - \left[u^{(i)}(y)^\top z\right]^2\right| \\ &= \left|\left[\frac{(\bar{x}-a^{(i)})^\top z}{\|\bar{x}-a^{(i)}\|_2}\right]^2 - \left[\frac{(y-a^{(i)})^\top z}{\|y-a^{(i)}\|_2}\right]^2\right| \\ &\le \left|\left[\frac{(\bar{x}-a^{(i)})^\top z}{\|\bar{x}-a^{(i)}\|_2}\right]^2 - \left[\frac{(\bar{x}-a^{(i)})^\top z}{\|y-a^{(i)}\|_2}\right]^2\right| + \left|\left[\frac{(\bar{x}-a^{(i)})^\top z}{\|y-a^{(i)}\|_2}\right]^2 - \left[\frac{(y-a^{(i)})^\top z}{\|y-a^{(i)}\|_2}\right]^2\right| \\ &\le \left|1 - \frac{\|\bar{x}-a^{(i)}\|_2^2}{\|y-a^{(i)}\|_2^2}\right| + \frac{\left|\left[(\bar{x}-a^{(i)}+\alpha v)^\top z\right]^2 - \left[(\bar{x}-a^{(i)})^\top z\right]^2\right|}{\|y-a^{(i)}\|_2^2} \\ &= \frac{\alpha^2 + \left|2\left[(\bar{x}-a^{(i)})^\top z\right]\cdot\left[\alpha v^\top z\right] + \left[\alpha v^\top z\right]^2\right|}{\|\bar{x}-a^{(i)}\|_2^2 + \alpha_i^2} \end{aligned}$$

where we used that $y = \bar{x} + \alpha v$ and $\|y - a^{(i)}\|_2^2 = \alpha^2 + \|\bar{x} - a^{(i)}\|_2^2$ (since $v \perp (\bar{x} - a^{(i)})$). Now we know that $\alpha^2 \le \frac{\epsilon^2}{t^2}$ and therefore, by Young's inequality and Cauchy Schwarz we have that for all $\gamma > 0$

$$\begin{aligned} \left|\left[u^{(i)}(x)^\top z\right]^2 - \left[u^{(i)}(y)z\right]^2\right| &\le \frac{2\alpha^2 + 2\left|\left[(\bar{x}-a^{(i)})^\top z\right]\cdot\left[\alpha v^\top z\right]\right|}{\|\bar{x}-a^{(i)}\|_2^2 + \alpha^2} \\ &\le \frac{2\alpha^2 + \gamma\left[(\bar{x}-a^{(i)})^\top z\right]^2 + \gamma^{-1}\alpha^2\left[v^\top z\right]^2}{\|\bar{x}-a^{(i)}\|_2^2 + \alpha^2} \\ &\le \frac{\alpha^2\left(2 + \gamma^{-1}\left(v^\top z\right)^2\right)}{\|\bar{x}-a^{(i)}\|_2^2 + \alpha^2} + \gamma\left[(u^{(i)}(x))^\top z\right]^2 \\ &\le \frac{\epsilon^2}{t^2\|\bar{x}-a^{(i)}\|_2^2 + \epsilon^2}\left(2 + \frac{1}{\gamma}\left(v^\top z\right)^2\right) + \gamma\left[(u^{(i)}(x))^\top z\right]^2 . \qquad \text{(A.1)} \end{aligned}$$

Note that

$$\begin{aligned} t^2\|\bar{x}-a^{(i)}\|_2^2 &= t^2\left(\|x-a^{(i)}\|_2^2 + 2\beta(x-a^{(i)})^\top u^{(i)}(x) + \beta^2\right) = \left(t\|x-a^{(i)}\|_2 + t\beta\right)^2 \\ &\ge \left(\max\left\{t\|x-a^{(i)}\|_2 - \epsilon, 0\right\}\right)^2 . \end{aligned}$$

Now, we separate the proof into two cases depending if $t\|x - a^{(i)}\|_2 \ge 2\epsilon^{1/2}\sqrt{g_t^{(i)}(x)}$.

If $t\|x - a^{(i)}\|_2 \ge 2\epsilon^{1/3}\sqrt{g_t^{(i)}(x)}$ then since $\epsilon \le \frac{1}{20}$ we have that

$$t\|x-a^{(i)}\|_2 \ge \left(\frac{t\|x-a^{(i)}\|_2}{\sqrt{g_t^{(i)}(x)}}\right)^2 \ge 4\epsilon^{2/3}.$$

and $t\|y - a^{(i)}\| \geq \epsilon$, justifying our assumption that $u^{(i)}(x) \neq 0$ and $u^{(i)}(y) \neq 0$. Furthermore, this implies that

$$t^2\|\bar{x} - a^{(i)}\|_2^2 \geq \left(\frac{3}{4}\right)^2 t^2\|x - a^{(i)}\|_2^2 \geq 2\epsilon^{2/3} g_t^{(i)}(x).$$

and therefore letting $\gamma = \frac{\epsilon^{2/3}}{g_t^{(i)}(x)}$ yields

$$\begin{aligned}
\left|\left[u_t^{(i)}(x)^\top z\right]^2 - \left[u_t^{(i)}(y) z\right]^2\right| &\leq \frac{\epsilon^{4/3}}{2g_t^{(i)}(x)}\left(2 + \frac{g_t^{(i)}(x)}{\epsilon^{2/3}}\left[v^\top z\right]^2\right) + \frac{\epsilon^{2/3}}{g_t^{(i)}(x)}\left[(u^{(i)}(x))^\top z\right]^2 \\
&\leq \frac{\epsilon^{2/3}}{2}\left[v^\top z\right]^2 + \frac{\epsilon^{4/3}}{g_t^{(i)}(x)} + \frac{\epsilon^{2/3}}{g_t^{(i)}(x)}\left[(u^{(i)}(x))^\top z\right]^2 \\
&\leq \frac{\epsilon^{2/3}}{2}\left[v^\top z\right]^2 + \frac{3}{2}\frac{\epsilon^{2/3}}{g_t^{(i)}(x)}.
\end{aligned}$$

Since $v \perp u^{(i)}(x)$ and v, z are unit vectors, both $\left[v^\top z\right]^2$ and $\frac{1}{g_t^{(i)}(x)}$ are less than

$$z^\top\left[\mathbf{I} - \left(1 - \frac{1}{g_t^{(i)}(x)}\right) u^{(i)}(y)(u^{(i)}(y))^\top\right] z.$$

Therefore, we have

$$\begin{aligned}
\left|\left[u_t^{(i)}(x)^\top z\right]^2 - \left[u_t^{(i)}(y) z\right]^2\right| &\leq 2\epsilon^{2/3} z^\top\left[\mathbf{I} - \left(1 - \frac{1}{g_t^{(i)}(x)}\right) u^{(i)}(y)(u^{(i)}(y))^\top\right] z \\
&= 2\epsilon^{2/3}\left(\frac{1 + g_t^{(i)}(x)}{t^2}\right)\|z\|^2_{\nabla^2 f_t^{(i)}(x)}
\end{aligned}$$

and therefore if we let

$$\mathbf{H} \overset{\text{def}}{=} \frac{t^2}{1 + g_t^{(i)}(x)}\left(\mathbf{I} - \left(1 - \frac{1}{g_t^{(i)}(x)}\right) u^{(i)}(y)(u^{(i)}(y))^\top\right),$$

we see that for unit vectors z,

$$\left|z^\top\left(\mathbf{H} - \nabla^2 f_t^{(i)}(x)\right) z\right| \leq 2\epsilon^{2/3}\|z\|^2_{\nabla^2 f_t^{(i)}(x)}$$

Otherwise, $t\|x - a^{(i)}\|_2 < 2\epsilon^{1/3}\sqrt{g_t^{(i)}(x)}$ and therefore

$$g_t^{(i)}(x)^2 = 1 + t^2\|x - a^{(i)}\|_2^2 \leq 1 + 4\epsilon^{2/3} g_t^{(i)}(x)$$

Therefore, we have

$$g_t^{(i)}(x) \quad \leq \quad \frac{4\epsilon^{2/3} + \sqrt{(4\epsilon^{2/3})^2 + 4}}{2} \leq 1 + 4\epsilon^{2/3}.$$

Therefore independent of (A.1) and the assumption that $u^{(i)}(x) \neq 0$ and $u^{(i)}(y) \neq 0$ we have

$$\frac{1}{1 + 4\epsilon^{2/3}}\mathbf{H} \preceq \frac{t^2}{(1 + g_t^{(i)}(x))g_t^{(i)}(x)}\mathbf{I} \preceq \nabla^2 f_t^{(i)}(x) \preceq \frac{t^2}{(1 + g_t^{(i)}(x))}\mathbf{I} \preceq \left(1 + 4\epsilon^{2/3}\right)\mathbf{H}.$$

In either case, we have that

$$\left| z^\top \left(\mathbf{H} - \nabla^2 f_t^{(i)}(x) \right) z \right| \le 4\epsilon^{2/3} \|z\|^2_{\nabla^2 f_t^{(i)}(x)}.$$

Now, we note that $\|x-y\|_2 \le \frac{\epsilon}{t} \le \epsilon \cdot \frac{g_t^{(i)}(x)}{t}$. Therefore, by Lemma 15 we have that

$$(1-\epsilon) g_t^{(i)}(x) \le g_t^{(i)}(y) \le (1+\epsilon) g_t^{(i)}(x)$$

Therefore, we have

$$\frac{1-4\epsilon^{2/3}}{(1+\epsilon)^2} \nabla^2 f_t^{(i)}(x) \preceq \frac{1}{(1+\epsilon)^2}\mathbf{H} \preceq \nabla^2 f_t^{(i)}(y) \preceq \frac{1}{(1-\epsilon)^2}\mathbf{H} \preceq \frac{1+4\epsilon^{2/3}}{(1-\epsilon)^2} \nabla^2 f_t^{(i)}(x)$$

Since $\epsilon < \frac{1}{20}$, the result follows. □

Consequently, so long as we have a point within a $O(\frac{1}{t})$ sized Euclidean ball of some x_t, Newton's method (or an appropriately transformed first order method) within the ball will converge quickly.

A.3 How Much Does the Hessian Change Along the Path?

Lemma 2. *For all $t \ge 0$ and $i \in [n]$ the following hold*

$$\left\| \frac{d}{dt} x_t \right\|_2 \le \frac{1}{t^2} \bar{g}_t(x_t) \quad , \quad \left| \frac{d}{dt} g_t^{(i)}(x_t) \right| \le \frac{1}{t} \left(g_t^{(i)}(x_t) + \bar{g}_t \right) \quad , \textit{ and } \quad \left| \frac{d}{dt} w_t \right| \le \frac{2}{t} w_t$$

Consequently, for all $t' \ge t$ we have that $\left(\frac{t}{t'}\right)^2 w_t \le w_{t'} \le \left(\frac{t'}{t}\right)^2 w_t$.

Proof of Lemma 2. From Lemma 13 we know that

$$\frac{d}{dt} x_t = -\left(\nabla^2 f_t(x_t)\right)^{-1} \sum_{i\in[n]} \frac{t(x_t - a^{(i)})}{(1+g_t^{(i)}(x_t)) g_t^{(i)}(x_t)}$$

and by Lemma 14 we know that

$$\nabla^2 f_t(x_t) \succeq \sum_{i\in[n]} \frac{t^2}{(1+g_t^{(i)}(x_t)) g_t^{(i)}(x_t)} \mathbf{I} = \frac{t^2}{\bar{g}_t(x_t)} \sum_{i\in[n]} \frac{1}{1+g_t^{(i)}(x_t)} \mathbf{I}\,.$$

Using this fact and the fact that $t\|x_t - a^{(i)}\|_2 \le g_t^{(i)}$ we have

$$\begin{aligned}\left\| \frac{d}{dt} x_t \right\|_2 &= \left\| -\left(\nabla^2 f_t(x_t)\right)^{-1} \frac{d}{dt} \nabla f_t(x_t) \right\|_2 \\ &\le \left(\frac{t^2}{\bar{g}_t(x_t)} \sum_{i\in[n]} \frac{1}{1+g_t^{(i)}(x_t)} \right)^{-1} \sum_{i\in[n]} \left\| \frac{t(x_t - a^{(i)})}{g_t^{(i)}(x_t)(1+g_t^{(i)}(x_t))} \right\|_2 \le \frac{\bar{g}_t(x_t)}{t^2}\,.\end{aligned}$$

Next, we have

$$\begin{aligned}\frac{d}{dt} g_t^{(i)}(x_t) &= \frac{d}{dt}\left(1 + t^2\|x_t - a^{(i)}\|_2^2\right)^{\frac{1}{2}} \\ &= \frac{1}{2} \cdot g_t^{(i)}(x_t)^{-1} \left(2t\|x_t - a^{(i)}\|_2^2 + 2t^2 (x_t - a^{(i)})^\top \frac{d}{dt} x_t \right)\end{aligned}$$

which by Cauchy Schwarz and that $t\|x_t - a^{(i)}\|_2 \leq g_t^{(i)}(x_t)$ yields the second equation. Furthermore,

$$\left|\frac{d}{dt}w_t\right| = \left|\frac{d}{dt}\sum_{i\in[n]}\frac{1}{1+g_t^{(i)}(x_t)}\right| \leq \sum_{i\in[n]}\left|\frac{d}{dt}\frac{1}{1+g_t^{(i)}(x_t)}\right| = \sum_{i\in[n]}\left|\frac{1}{(1+g_t^{(i)}(x_t))^2}\frac{d}{dt}g_t^{(i)}(x_t)\right|$$
$$\leq \frac{1}{t}\sum_{i\in[n]}\frac{g_t^{(i)}(x_t)+\bar{g}_t}{(1+g_t^{(i)}(x_t))^2} \leq 2\frac{w_t}{t}$$

which yields the third equation. Therefore, we have that

$$|\ln w_{t'} - \ln w_t| = \left|\int_t^{t'}\frac{\frac{d}{d\alpha}w_\alpha}{w_\alpha}d\alpha\right| \leq \int_t^{t'}\frac{(2\frac{w_\alpha}{\alpha})}{w_\alpha}d\alpha = 2\int_t^{t'}\frac{1}{\alpha}d\alpha = \ln\left(\frac{t'}{t}\right)^2.$$

Exponentiating the above inequality yields the final inequality. □

Lemma 3. *For all $t \geq 0$ we have*

$$-12\cdot t\cdot w_t\mathbf{I} \preceq \frac{d}{dt}\left[\nabla^2 f_t(x_t)\right] \preceq 12\cdot t\cdot w_t\mathbf{I} \tag{3.1}$$

and therefore for all $\beta \in [0, \frac{1}{8}]$

$$\nabla^2 f(x_t) - 15\beta t^2 w_t\mathbf{I} \preceq \nabla^2 f(x_{t(1+\beta)}) \preceq \nabla^2 f(x_t) + 15\beta t^2 w_t\mathbf{I}\,. \tag{3.2}$$

Proof of Lemma 3. Let

$$\mathbf{A}_t^{(i)} \stackrel{\text{def}}{=} \frac{t^2(x_t - a^{(i)})(x_t - a^{(i)})^\top}{(1+g_t^{(i)})g_t^{(i)}}$$

and recall that $\nabla^2 f_t(x_t) = \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}}\left(\mathbf{I} - \mathbf{A}_t^{(i)}\right)$. Consequently

$$\frac{d}{dt}\nabla^2 f_t(x_t) = \frac{d}{dt}\left(\sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}}\left(\mathbf{I} - \mathbf{A}_t^{(i)}\right)\right)$$
$$= 2t\left(\frac{1}{t^2}\right)\nabla^2 f_t(x_t) + t^2\sum_{i\in[n]}\frac{-\frac{d}{dt}g_t^{(i)}}{(1+g_t^{(i)})^2}\left(\mathbf{I} - \mathbf{A}_t^{(i)}\right) - \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}}\frac{d}{dt}\mathbf{A}_t^{(i)}$$

Now, since $\mathbf{0} \preceq \mathbf{A}_t^{(i)} \preceq \mathbf{I}$ we have $0 \preceq \nabla^2 f_t(x_t) \preceq t^2 w_t\mathbf{I}$. For all unit vectors v, using Lemma 2 yields

$$\left|v^\top\left(\frac{d}{dt}\nabla^2 f_t(x_t)\right)v\right| \leq 2t\cdot w_t\cdot\|v\|_2^2 + t^2\sum_{i\in[n]}\frac{\left|\frac{d}{dt}g_t^{(i)}\right|}{(1+g_t^{(i)})^2}\|v\|_2^2 + \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}}\left|v^\top\left(\frac{d}{dt}\mathbf{A}_t^{(i)}\right)v\right|$$
$$\leq 4t\cdot w_t + \sum_{i\in[n]}\frac{t^2}{1+g_t^{(i)}}\left|v^\top\left(\frac{d}{dt}\mathbf{A}_t^{(i)}\right)v\right|.$$

Next

$$\frac{d}{dt}\mathbf{A}_t^{(i)} = 2t\left(\frac{1}{t^2}\right)\mathbf{A}_t^{(i)} - \left(\frac{t}{(1+g_t^{(i)})g_t^{(i)}}\right)^2\left[(1+g_t^{(i)})\frac{d}{dt}g_t^{(i)} + g_t^{(i)}\frac{d}{dt}g_t^{(i)}\right](x_t - a^{(i)})(x_t - a^{(i)})^\top$$
$$+ \frac{t^2}{(1+g_t^{(i)})g_t^{(i)}}\left[(x_t - a^{(i)})\left(\frac{d}{dt}x_t\right)^\top + \left(\frac{d}{dt}x_t\right)(x_t - a^{(i)})^\top\right],$$

and therefore by Lemma 2 and the fact that $t\|x_t - a^{(i)}\|_2 \le g_t^{(i)}$ we have

$$\begin{aligned}\left|v^\top\left(\frac{d}{dt}\mathbf{A}_t^{(i)}\right)v\right| &\le \left(\frac{2}{t} + \frac{2t^2\left|\frac{d}{dt}g_t^{(i)}\right|}{(1+g_t^{(i)})(g_t^{(i)})^2}\|x_t - a^{(i)}\|_2^2 + \frac{2t^2\|x_t - a^{(i)}\|_2\|\frac{d}{dt}x_t\|_2}{(1+g_t^{(i)})g_t^{(i)}}\right)\|v\|_2^2 \\ &\le \frac{2}{t} + \frac{2}{t}\cdot\frac{g_t^{(i)} + \bar{g}_t}{1+g_t^{(i)}} + \frac{2}{t}\cdot\frac{\bar{g}_t}{1+g_t^{(i)}} \le \frac{4}{t} + \frac{4}{t}\frac{\bar{g}_t}{1+g_t^{(i)}}\,.\end{aligned}$$

Consequently, we have

$$\left|v^\top\left(\frac{d}{dt}\nabla^2 f_t(x_t)\right)v\right| \le 8t\cdot w_t + 4t\sum_{i\in[n]}\frac{\bar{g}_t}{(1+g_t^{(i)})^2} \le 12t\cdot w_t$$

which completes the proof of (3.1). To prove (3.2), let v be any unit vector and note that

$$\begin{aligned}\left|v^\top\left(\nabla^2 f_{t(1+\beta)}(x) - \nabla^2 f_t(x)\right)v\right| &= \left|\int_t^{t(1+\beta)} v^\top\frac{d}{d\alpha}\left[\nabla^2 f_\alpha(x_\alpha)\right]v\cdot d\alpha\right| \le 12\int_t^{t(1+\beta)}\alpha\cdot w_\alpha d\alpha \\ &\le 12\int_t^{t(1+\beta)}\alpha\left(\frac{\alpha}{t}\right)^2 w_t d\alpha \le \frac{12}{t^2}\left(\frac{1}{4}\left[t(1+\beta)\right]^4 - \frac{1}{4}t^4\right)w_t \\ &= 3t^2\left[(1+\beta)^4 - 1\right]w_t \le 15t^2\beta w_t\end{aligned}$$

where we used Lemma 3 and $0 \le \beta \le \frac{1}{8}$ at the last line. □

A.4 Where is the next Optimal Point?

Lemma 16. *For all t, we have*

$$\frac{1}{2}\left[t^2\cdot w_t\mathbf{I} - (t^2\cdot w_t - \mu_t)v_t v_t^\top\right] \preceq \nabla^2 f_t(x_t) \preceq t^2\cdot w_t\mathbf{I} - (t^2\cdot w_t - \mu_t)v_t v_t^\top.$$

Proof of Lemma 16. This follows immediately from Lemma 14, regarding the hessian of the penalized objective function, and Lemma 26, regarding the sum of PSD matrices expressed as the identity matrix minus a rank 1 matrix. □

Lemma 5 (The Central Path is Almost Straight)**.** *For all $t \ge 0$, $\beta \in [0, \frac{1}{600}]$, and any unit vector y with $|\langle y, v_t\rangle| \le \frac{1}{t^2\cdot\kappa}$ where $\kappa = \max_{\delta\in[t,(1+\beta)t]}\frac{w_\delta}{\mu_\delta}$, we have $y^\top(x_{(1+\beta)t} - x_t) \le \frac{6\beta}{t}$.*

Proof of Lemma 5. Clearly

$$\begin{aligned}y^\top(x_{(1+\beta)t} - x_t) &= \int_t^{(1+\beta)t} y^\top\frac{d}{d\alpha}x_\alpha d\alpha \le \int_\beta^{(1+\beta)t}\left|y^\top\frac{d}{d\alpha}x_\alpha\right|d\alpha \\ &\le \int_t^{(1+\beta)t}\left|y^\top\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1}\sum_{i\in[n]}\frac{\alpha}{(1+g_\alpha^{(i)})g_\alpha^{(i)}}(x_\alpha - a^{(i)})\right|d\alpha \\ &\le \int_t^{(1+\beta)t}\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1}y\|_2\cdot\left\|\sum_{i\in[n]}\frac{\alpha}{(1+g_\alpha^{(i)})g_\alpha^{(i)}}(x_\alpha - a^{(i)})\right\|_2 d\alpha\end{aligned}$$

Now since clearly $\alpha\|x_\alpha - a^{(i)}\|_2 \leq g_\alpha^{(i)}$, invoking Lemma 2 yields that

$$\left\|\sum_{i\in[n]} \frac{\alpha(x_\alpha - a^{(i)})}{(1+g_\alpha^{(i)})g_\alpha^{(i)}}\right\|_2 \leq \sum_{i\in[n]} \frac{1}{1+g_\alpha^{(i)}} = w_\alpha \leq \left(\frac{\alpha}{t}\right)^2 w_t \,.$$

Now by invoking Lemma 3 and the Lemma 16, we have that

$$\nabla^2 f_\alpha(x_\alpha) \succeq \nabla^2 f_t(x_t) - 15\beta t^2 w_t \mathbf{I} \succeq \frac{1}{2}\left[t^2 \cdot w_t \mathbf{I} - (t^2 \cdot w_t - \mu_t) v_t v_t^\top\right] - 15\beta t^2 w_t \mathbf{I}.$$

For notational convenience let $\mathbf{H}_t \stackrel{\mathrm{def}}{=} \nabla^2 f_t(x_t)$ for all $t > 0$. Then Lemma 3 shows that $\mathbf{H}_\alpha = \mathbf{H}_t + \Delta_\alpha$ where $\|\Delta_\alpha\|_2 \leq 15\beta t^2 w_t$. Now, we note that

$$\mathbf{H}_\alpha^2 = \mathbf{H}_t^2 + \Delta_\alpha \mathbf{H}_t + \mathbf{H}_t \Delta_\alpha + \Delta_\alpha^2 \,.$$

Therefore, we have

$$\begin{aligned}\|\mathbf{H}_\alpha^2 - \mathbf{H}_t^2\|_2 &\leq \|\Delta_\alpha \mathbf{H}_t\|_2 + \|\mathbf{H}_t \Delta_\alpha\|_2 + \|\Delta_\alpha^2\|_2 \\ &\leq 2\|\Delta\|_2 \|\mathbf{H}_t\|_2 + \|\Delta\|_2^2 \leq 40\beta t^4 w_t^2 \,.\end{aligned}$$

Let S be the subspace orthogonal to v_t. Then, Lemma 16 shows that $\mathbf{H}_t \succeq \frac{1}{2} t^2 w_t \mathbf{I}$ on S and hence $\mathbf{H}_t^2 \succeq \frac{1}{4} t^4 w_t^2 \mathbf{I}$ on S.[4] Since $\|\mathbf{H}_\alpha^2 - \mathbf{H}_t^2\|_2 \leq 40\beta t^4 w_t^2$, we have that

$$\mathbf{H}_\alpha^2 \succeq \left(\frac{1}{4} t^4 w_t^2 - 40\beta t^4 w_t^2\right) \mathbf{I} \text{ on } S$$

and hence

$$\mathbf{H}_\alpha^{-2} \preceq \left(\frac{1}{4} t^4 w_t^2 - 40\beta t^4 w_t^2\right)^{-1} \mathbf{I} \text{ on } S.$$

Therefore, for any $z \in S$, we have

$$\left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} z\right\|_2 = \left\|\mathbf{H}_\alpha^{-1} z\right\|_2 \leq \frac{\|z\|_2}{\sqrt{\frac{1}{4} t^4 w_t^2 - 40\beta t^4 w_t^2}}.$$

Now, we split $y = z + \langle y, v_t\rangle v_t$ where $z \in S$. Then, we have that

$$\begin{aligned}\left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} y\right\|_2 &\leq \left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} z\right\|_2 + |\langle y, v_t\rangle| \left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} v_t\right\|_2 \\ &\leq \frac{1}{\sqrt{\frac{1}{4} t^4 w_t^2 - 40\beta t^4 w_t^2}} + \frac{1}{t^2 \cdot \kappa} \left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} v_t\right\|_2 .\end{aligned}$$

Note that, we also know that $\lambda_{\min}(\nabla^2 f_\alpha(x_\alpha)) \geq \mu_\alpha$ and hence $\lambda_{\max}(\nabla^2 f_\alpha(x_\alpha)^{-2}) \leq \mu_\alpha^{-2}$. Therefore, we have

$$\left\|\left(\nabla^2 f_\alpha(x_\alpha)\right)^{-1} y\right\|_2 \leq \frac{1}{t^2 w_t \sqrt{\frac{1}{4} - 40\beta}} + \frac{1}{t^2} \frac{\mu_\alpha}{w_\alpha} \frac{1}{\mu_\alpha} \leq \frac{1}{t^2 w_t}\left(2 + \frac{1}{\sqrt{\frac{1}{4} - 40\beta}}\right) \leq \frac{5}{t^2 w_t} \,.$$

Combining these and using that $\beta \in [0, 1/600]$ yields that

$$\begin{aligned}y^\top(x_{(1+\beta)t} - x_t) &\leq \int_t^{(1+\beta)t} \frac{5}{t^2 w_t} \left(\frac{\alpha}{t}\right)^2 w_t d_\alpha \leq \frac{5}{t^4}\left(\frac{1}{3}(1+\beta)^3 t^3 - \frac{1}{3} t^3\right) \\ &\leq \frac{5}{3t}\left[(1+\beta)^3 - 1\right] \leq \frac{6\beta}{t}.\end{aligned}$$

□

[4] By $\mathbf{A} \preceq \mathbf{B}$ on S we mean that for all $x \in S$ we have $x^\top \mathbf{A} x \leq x^\top \mathbf{B} x$. The meaning of $\mathbf{A} \succeq \mathbf{B}$ on S is analagous.

A.5 Where is the End?

Lemma 6. *$f(x_t) - f(x_*) \leq \frac{2n}{t}$ for all $t > 0$.*

Proof of Lemma 6. Clearly, $\nabla f_t(x_t) = 0$ by definition of x_t. Consequently $\frac{1}{t}\nabla f_t(x_t)^\top (x_t - x_*) = 0$ and using Lemma 13 to give the formula for $\nabla f_t(x_t)$ yields

$$0 = \sum_{i\in[n]} \frac{t(x_t - a^{(i)})^\top (x_t - x_*)}{1 + g_t^{(i)}(x)} = \sum_{i\in[n]} \frac{t\|x_t - a^{(i)}\|_2^2 + t(x_t - a^{(i)})^\top (a^{(i)} - x_*)}{1 + g_t^{(i)}(x_t)} .$$

Therefore, by Cauchy Schwarz and the fact that $t\|x_t - a^{(i)}\|_2 \leq g_t^{(i)}(x_t) \leq 1 + g_t^{(i)}$

$$\sum_{i\in[n]} \frac{t(x_t - a^{(i)})^\top (a^{(i)} - x_*)}{1 + g_t^{(i)}(x_t)} \geq - \sum_{i\in[n]} \frac{t\|x_t - a^{(i)}\|_2 \|a^{(i)} - x_*\|_2}{1 + g_t^{(i)}(x_t)} \geq -f(x_*) .$$

Furthermore, since $1 + g_t^{(i)}(x_t) \leq 2 + t\|x_t - a^{(i)}\|_2$ we have

$$\sum_{i\in[n]} \frac{t\|x_t - a^{(i)}\|_2^2}{1 + g_t^{(i)}(x_t)} \geq \sum_{i\in[n]} \|x_t - a^{(i)}\|_2 - \sum_{i\in[n]} \frac{2\|x_t - a^{(i)}\|_2}{1 + g_t^{(i)}(x_t)} \geq f(x_t) - \frac{2n}{t} .$$

Combining yields the result. □

A.6 Simple Lemmas

Here we provide various small technical Lemmas that we will use to bound the accuracy with which we need to carry out various operations in our algorithm. Here we use some notation from Section 4 to simplify our bounds and make them more readily applied.

Lemma 17. *For any x, we have that $\|x - x_t\|_2 \leq f(x)$.*

Proof of Lemma 17. Since $\sum_{i\in[n]} \|x - a^{(i)}\|_2 = f(x)$, we have that $\|x - a^{(i)}\|_2 \leq f(x)$ for all $i \in [n]$. Since $\nabla f(x_t) = 0$ by Lemma 13 we see that x_t is a convex combination of the $a^{(i)}$ and therefore $\|x - x_t\|_2 \leq f(x)$ by convexity. □

Lemma 18. *$x^{(0)} = \frac{1}{n}\sum_{i\in[n]} a^{(i)}$ is a 2-approximate geometric median, i.e. $\tilde{f}_* \leq 2 \cdot f(x_*)$.*

Proof. For all $x \in \mathbb{R}^d$ we have

$$\|x^{(0)} - x\|_2 = \left\| \frac{1}{n}\sum_{i\in[n]} a^{(i)} - \frac{1}{n}\sum_{i\in[n]} x \right\|_2 \leq \frac{1}{n}\sum_{i\in[n]} \|a^{(i)} - x\|_2 \leq \frac{f(x)}{n} .$$

Consequently,

$$f(x^{(0)}) \leq \sum_{i\in[n]} \|x^{(0)} - a^{(i)}\|_2 \leq \sum_{i\in[n]} \left(\|x^{(0)} - x_*\|_2 + \|x_* - a^{(i)}\|_2 \right) \leq 2 \cdot f(x_*)$$

□

Lemma 19. *For all $t \geq 0$, we have*

$$1 \leq \frac{t^2 \cdot w_t(x)}{\mu_t(x)} \leq \bar{g}_t(x) \leq \max_{i \in [n]} g_t^{(i)}(x) \leq 1 + t \cdot f(x)\,.$$

In particular, if $t \leq \frac{2n}{\tilde{\epsilon}_* \cdot f(x_*)}$, we have that

$$g_t^{(i)} \leq \frac{3n}{\tilde{\epsilon}_*} + tn\|x - x_t\|_2.$$

Proof of Lemma 19. The first claim $1 \leq \frac{t^2 \cdot w_t(x)}{\mu_t(x)} \leq \bar{g}_t(x)$, follows from $\mu_t(x) \geq \sum_{i \in [n]} \frac{t^2}{g_t^{(i)}(x)(1+g_t^{(i)}(x))}$ and the fact that the largest eigenvalue of $\nabla^2 f_t(x)$ is at most $t^2 \cdot w_t(x)$. The second follows from the fact that $\bar{g}_t(x)$ is a weighted harmonic mean of $g_t^{(i)}(x)$ and therefore

$$\bar{g}_t(x) \leq \max_{i \in [n]} g_t^{(i)}(x) \leq 1 + t \cdot \max_{i \in [n]} \|x - a^{(i)}\|_2 \leq 1 + t \cdot f(x)\,.$$

For the final inequality, we use the fact that $f(x) \leq f(x_t) + n\|x - x_t\|_2$ and the fact that $f(x_t) \leq f(x_*) + \frac{2n}{t}$ by Lemma 6 and get

$$g_t^{(i)} \leq 1 + t\left(f(x_*) + \frac{2n}{t} + n\|x - x_t\|_2\right) \leq \frac{3n}{\tilde{\epsilon}_*} + tn\|x - x_t\|_2\,.$$

□

Lemma 20. *For all $x \in \mathbb{R}^d$ and $t > 0$, we have*

$$\frac{n}{2}\left(\frac{\|x - x_t\|_2}{\frac{3n}{t \cdot \tilde{\epsilon}_*} + n\|x - x_t\|_2}\right)^2 \leq f_t(x) - f_t(x_t) \leq \frac{nt^2}{2}\|x - x_t\|_2^2$$

Proof of Lemma 20. For the first inequality, note that $\nabla^2 f_t(x) \preceq \sum_{i \in [n]} \frac{t^2}{1+g_t^{(i)}(x)}\mathbf{I} \preceq n \cdot t^2 \mathbf{I}$. Consequently, if we let $n \cdot t^2 \mathbf{I} = \mathbf{H}$ in Lemma 30, we have that

$$f_t(x) - f_t(x_t) \leq \frac{1}{2}\|x - x_t\|_{\mathbf{H}}^2 \leq \frac{nt^2}{2}\|x - x_t\|_2^2\,.$$

For the second inequality, note that Lemma 14 and Lemma 19 yields that

$$\nabla^2 f_t(x) \succeq \sum_{i \in [n]} \frac{t^2}{(1 + g_t^{(i)}(x))g_t^{(i)}(x)}\mathbf{I} \succeq n\left(\frac{t}{\frac{3n}{\tilde{\epsilon}_*} + tn\|x - x_t\|_2}\right)^2 \mathbf{I}\,.$$

Consequently, applying 30 again yields the lower bound. □

B Nearly Linear Time Geometric Median (Proofs)

Here we provide proofs, algorithms, and technical lemmas from Section 4.

B.1 Eigenvector Computation and Hessian Approximation

Below we prove that the power method can be used to compute an ϵ-approximate top eigenvector of a symmetric PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with a non-zero eigenvalue gap $g = \frac{\lambda_1(\mathbf{A})-\lambda_2(\mathbf{A})}{\lambda_1(\mathbf{A})}$. While it is well know that this can be by applying $\mathbf{A}$ to a random initial vector $O(\frac{\alpha}{g}\log(\frac{d}{\epsilon}))$ times in the following theorem we provide a slightly less known refinement that the dimension d can be replaced with the stable rank of $\mathbf{A}$, $s = \sum_{i\in[d]} \frac{\lambda_i(\mathbf{A})}{\lambda_1(\mathbf{A})}$. We use this fact to avoid a dependence on d in our logarithmic factors.

```
Algorithm 5: PowerMethod($\mathbf{A}, k$)
  Input: symmetric PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and a number of iterations $k \geq 1$.
  Let $x \sim \mathcal{N}(0, \mathbf{I})$ be drawn from a $d$ dimensional normal distribution.
  Let $y = \mathbf{A}^k x$
  Output: $u = y/\|y\|_2$
```

Lemma 21 (Power Method)**.** *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be a symmetric PSD matrix , let* $g \overset{\text{def}}{=} \frac{\lambda_1(\mathbf{A})-\lambda_2(\mathbf{A})}{\lambda_1(\mathbf{A})}$, $s = \sum_{i\in d} \frac{\lambda_i(\mathbf{A})}{\lambda_1(\mathbf{A})}$, *and let* $\epsilon > 0$ *and* $k \geq \frac{\alpha}{g}\log(\frac{ns}{\epsilon})$ *for large enough constant* α*. In time* $O(\mathrm{nnz}(\mathbf{A}) \cdot \log(\frac{ns}{\epsilon}))$*, the algorithm* ***PowerMethod***$(\mathbf{A}, k)$ *outputs a vector* u *such that* $\langle v_1(\mathbf{A}), u\rangle^2 \geq 1-\epsilon$ *and* $u^\top \mathbf{A} u \geq (1-\epsilon)\lambda_1(\mathbf{A})$*with high probability in* n/ϵ*.*

Proof. We write $x = \sum_{i\in[d]} \alpha_i v_i(\mathbf{A})$. Then, we have

$$\langle v_1(\mathbf{A}), u\rangle^2 = \left\langle v_1(\mathbf{A}), \frac{\sum_{i\in[d]} \alpha_i \lambda_i(\mathbf{A})^k v_i(\mathbf{A})}{\sqrt{\sum_{i\in[d]} \alpha_i^2 \lambda_i(\mathbf{A})^{2k}}} \right\rangle^2 = \frac{\alpha_1^2}{\alpha_1^2 + \sum_{j\neq 1} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^{2k}} \geq 1 - \sum_{j\neq 1} \frac{\alpha_j^2}{\alpha_1^2}\left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^{2k}$$

Re arranging terms we have

$$\begin{aligned} 1 - \langle v_1(\mathbf{A}), u\rangle^2 &\leq \sum_{j\neq 1} \frac{\alpha_j^2}{\alpha_1^2}\left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)\left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^{2k-1} \leq \sum_{j\neq 1} \frac{\alpha_j^2}{\alpha_1^2}\left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)\left(\frac{\lambda_2(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^{2k-1} \\ &= \sum_{j\neq 1,} \frac{\alpha_j^2}{\alpha_1^2} \cdot \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right) \cdot (1-g)^{2k-1} \leq \sum_{j\neq 1} \frac{\alpha_j^2}{\alpha_1^2} \cdot \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right) \cdot \exp(-(2k-1)g) \end{aligned}$$

where we used that $\frac{\lambda_2}{\lambda_1} = 1 - g \leq e^{-g}$.

Now with high probability in n/ϵ we have that $\alpha_1^2 \geq \frac{1}{O(\mathsf{poly}(n/\epsilon))}$ by known properties of the chi-squared distribution. All that remains is to upper bound $\sum_{j\neq 1} \alpha_j^2 \cdot \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)$. To bound this consider $h(\alpha) \overset{\text{def}}{=} \sqrt{\sum_{j\neq 1} \alpha_j^2(\lambda_j(\mathbf{A})/\lambda_1(\mathbf{A}))}$. Note that

$$\|\nabla h(\alpha)\|_2 = \left\| \frac{\sum_{j\neq 1} \vec{1}_j \cdot \alpha_j \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A}}\right)}{\sqrt{\sum_{j\neq 1} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)}} \right\|_2 = \sqrt{\frac{\sum_{j\neq 1} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)^2}{\sum_{j\neq 1} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)}} \leq 1\,.$$

where $\vec{1}_j$ is the indicator vector for coordinate j. Consequently h is 1-Lipschitz and by Gaussian concentration for Lipschitz functions we know there are absolute constants C and c such that

$$\Pr\left[h(\alpha) \geq \mathbb{E}h(\alpha) + \lambda\right] \leq C\exp(-c\lambda^2)\,.$$

By the concavity of square root and the expected value of the chi-squared distribution we have

$$\mathbb{E}h(\alpha) \leq \sqrt{\mathbb{E}\sum_{j\neq i}\alpha_j^2 \cdot \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)} = \sqrt{\sum_{j\neq i}\left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right)} \leq \sqrt{s}\,.$$

Consequently, since $s \geq 1$ we have that $\Pr[h(\alpha) \geq (1+\lambda)\cdot\sqrt{s}] \leq C\exp(-c\cdot\lambda^2)$ for $\lambda \geq 1$ and that $\sum_{j\neq 1}\alpha_j \cdot \left(\frac{\lambda_j(\mathbf{A})}{\lambda_1(\mathbf{A})}\right) = O(ns/\epsilon)$ with high probability in n/ϵ. Since $k = \Omega(\frac{1}{g}\log(\frac{ns}{\epsilon}))$, we have $\langle v_1(\mathbf{A}), u\rangle^2 \geq 1-\epsilon$ with high probability in n/ϵ. Furthermore, this implies that

$$u^\top \mathbf{A} u = u^\top \left(\sum_{i\in[d]}\lambda_i(\mathbf{A})v_i(\mathbf{A})v_i(\mathbf{A})^\top\right) u \geq \lambda_1(\mathbf{A})\langle v_1(\mathbf{A}), u\rangle^2 \geq (1-\epsilon)\lambda_1(\mathbf{A})\,.$$

□

Lemma 7 (Computing Hessian Approximation)**.** *Let $x \in \mathbb{R}^d$, $t > 0$, and $\epsilon \in (0, \frac{1}{4})$. The algorithm* `ApproxMinEig`$(x,t,\epsilon)$ *outputs (λ, u) in $O(nd\log\frac{n}{\epsilon})$ time such that if $\mu_t(x) \leq \frac{1}{4}t^2 w_t(x)$ then $\langle v_t(x), u\rangle^2 \geq 1-\epsilon$ with high probability in n/ϵ. Furthermore, if $\epsilon \leq \left(\frac{\mu_t(x)}{8t^2\cdot w_t(x)}\right)^2$ then $\frac{1}{4}\mathbf{Q} \preceq \nabla^2 f_t(x) \preceq 4\mathbf{Q}$ with high probability in n/ϵ where $\mathbf{Q} \stackrel{\text{def}}{=} t^2\cdot w_t(x) - \left(t^2\cdot w_t(x) - \lambda\right)uu^\top$.*

Proof of Lemma 7. By Lemma 16 we know that $\frac{1}{2}\mathbf{Z} \preceq \nabla^2 f_t(x) \preceq \mathbf{Z}$ where

$$\mathbf{Z} = t^2\cdot w_t(x)\mathbf{I} - \left(t^2\cdot w_t(x) - \mu_t(x)\right)v_t(x)v_t(x)^\top.$$

Consequently, if $\mu_t(x) \leq \frac{1}{4}t^2 w_t(x)$, then for all unit vectors $z \perp v_t(x)$, we have that

$$z^\top\nabla^2 f_t(x) z \geq \frac{1}{2}z^\top\mathbf{Z}z \geq \frac{1}{2}t^2 w_t(x).$$

Since $\nabla^2 f_t(x) = t^2\cdot w_t(x) - \mathbf{A}$, for $\mathbf{A}$ in the definition of `ApproxMinEig` (Algorithm 2) this implies that $v_t(x)^\top\mathbf{A}v_t(x) \geq \frac{3}{4}t^2\cdot w_t(x)$ and $z^\top\mathbf{A}z \leq \frac{1}{2}t^2 w_t(x)$. Furthermore, we see that

$$\sum_{i\in[d]}\lambda_i(\mathbf{A}) = \operatorname{tr}(\mathbf{A}) = \sum_{i\in[n]}\frac{t^4\|x-a^{(i)}\|_2^2}{(1+g_t^{(i)}(x))^2 g_t^{(i)}(x)} \leq t^2\cdot w_t(x)$$

Therefore, in this case, $\mathbf{A}$ has a constant multiplicative gap between its top two eigenvectors and stable rank at most a constant (i.e. $g = \Omega(1)$ and $s = O(1)$ in Theorem 21). Consequently, by Theorem 21 we have $\langle v_t(x), u\rangle^2 \geq 1-\epsilon$.

For the second claim, we note that

$$t^2\cdot w_t(x) - \mu_t(x) \geq u^\top\mathbf{A}u \geq (1-\epsilon)\lambda_1(\mathbf{A}) = (1-\epsilon)(t^2\cdot w_t(x) - \mu_t(x))$$

Therefore, since $\lambda = u^\top\nabla^2 f_t(x)u = t^2\cdot w_t(x) - u^\top\mathbf{A}u$, we have

$$(1-\epsilon)\mu_t(x) - \epsilon\cdot t^2 w_t(x) \leq \lambda \leq \mu_t(x). \tag{B.1}$$

On the other hand, by Lemma 27, we have that

$$\sqrt{\epsilon}\mathbf{I} \preceq v_t(x)v_t(x)^\top - uu^\top \preceq \sqrt{\epsilon}\mathbf{I}. \tag{B.2}$$

Combining (B.1) and (B.2), we have $\frac{1}{2}\mathbf{Z} \preceq \mathbf{Q} \preceq 2\mathbf{Z}$ if $\epsilon \leq \left(\frac{\mu_t(x)}{8t^2\cdot w_t(x)}\right)^2$ and $\frac{1}{4}\mathbf{Q} \preceq \nabla^2 f_t(x) \preceq 4\mathbf{Q}$ follows.

On the other hand, when $\mu_t(x) > \frac{1}{4}t^2 w_t(x)$. It is the case that $\frac{1}{4}t^2\cdot w_t(x)\mathbf{I} \preceq \nabla^2 f_t(x) \preceq t^2\cdot w_t(x)\mathbf{I}$ and $\frac{1}{4}t^2\cdot w_t(x)\mathbf{I} \preceq \mathbf{Q} \preceq t^2\cdot w_t(x)\mathbf{I}$ again yielding $\frac{1}{4}\mathbf{Q} \preceq \nabla^2 f_t(x) \preceq 4\mathbf{Q}$. □

Lemma 8. *Let $(\lambda, u) = \texttt{ApproxMinEig}(x, t, \epsilon_v)$ for $\epsilon_v < \frac{1}{8}$ and $\|x - x_t\|_2 \leq \frac{\epsilon_c}{t}$ for $\epsilon_c \leq (\frac{\epsilon_v}{36})^{\frac{3}{2}}$. If $\mu_t \leq \frac{1}{4}t^2 \cdot w_t$ then with high probability in n/ϵ_v for all unit vectors $y \perp u$, we have $\langle y, v_t\rangle^2 \leq 8\epsilon_v$.*

Proof of Lemma 8. By Lemma 7 we know that $\langle v_t(x), u\rangle^2 \geq 1 - \epsilon_v$. Since clearly $\|x - x_t\|_2 \leq \frac{1}{20t}$, by assumption, Lemma 1 shows

$$(1 - 6\epsilon_c^{2/3})\nabla^2 f_t(x_t) \preceq \nabla^2 f_t(x) \preceq (1 + 6\epsilon_c^{2/3})\nabla^2 f_t(x_t).$$

Furthermore, since $\mu_t \leq \frac{1}{4}t^2 \cdot w_t$, as in Lemma 7 we know that the largest eigenvalue of $\mathbf{A}$ defined in $\texttt{ApproxMinEig}(x, t, \epsilon)$ is at least $\frac{3}{4}t^2 \cdot w_t$ while the second largest eigenvalue is at most $\frac{1}{2}t^2 \cdot w_t$. Consequently, the eigenvalue gap, g, defined in Lemma 28 is at least $\frac{1}{3}$ and this lemma shows that $\langle v_t, u\rangle^2 \geq 1 - 36\epsilon_c^{2/3} \geq 1 - \epsilon_v$. Consequently, by Lemma 29, we have that $\langle u, v_t\rangle^2 \geq 1 - 4\epsilon_v$.

To prove the final claim, we write $u = \alpha v_t + \beta w$ for an unit vector $w \perp v_t$. Since $y \perp u$, we have that $0 = \alpha\langle v_t, y\rangle + \beta\langle w, y\rangle$. Then, either $\langle v_t, y\rangle = 0$ and the result follows or $\alpha^2\langle v_t, y\rangle^2 = \beta^2\langle w, y\rangle^2$ and since $\alpha^2 + \beta^2 = 1$, we have

$$\langle v_t, y\rangle^2 \leq \frac{\beta^2\langle w, y\rangle^2}{\alpha^2} \leq \frac{1 - \alpha^2}{\alpha^2} \leq 2(1 - \alpha^2) \leq 8\epsilon_v$$

where in the last line we used that $\alpha^2 \geq 1 - 4\epsilon_v > \frac{1}{2}$ since $\epsilon_v \leq \frac{1}{8}$. □

Lemma 9. *Suppose $\mu_t \geq \frac{1}{4}t^2 \cdot w_t$ and let $t' \in [t, (1 + \frac{1}{600})t]$ then $\|x_{t'} - x_t\|_2 \leq \frac{1}{100t}$.*

Proof of Lemma 9. Note that $t' = (1 + \beta)t$ where $\beta \in [0, \frac{1}{600}]$. Since $\frac{1}{4}t^2 \cdot w_t\mathbf{I} \preceq \mu_t\mathbf{I} \preceq \nabla^2 f(x_t)$ applying Lemma 3 then yields that for all $s \in [t, t']$

$$\nabla^2 f(x_s) \succeq \nabla^2 f(x_t) - 15\beta t^2 w_t\mathbf{I} \succeq \left(\frac{1}{4} - 15\beta\right) t^2 \cdot w_t\mathbf{I} \succeq \frac{t^2 \cdot w_t}{5}\mathbf{I}\,.$$

Consequently, by Lemma 13, the fact that $t\|x_t - a^{(i)}\|_2 \leq g_t^{(i)}$, and Lemma 2 we have

$$\begin{aligned}
\|x_{t'} - x_t\|_2 &\leq \int_t^{t'} \left\|\frac{d}{ds}x_s\right\|_2 d_s = \int_t^{t'} \left\|\left(\nabla^2 f_s(x_s)\right)^{-1} \sum_{i\in[n]} \frac{s}{(1 + g_s^{(i)})g_s^{(i)}}(x_s - a^{(i)})\right\|_2 d_s \\
&\leq \int_t^{t'} \frac{5}{t^2 \cdot w_t} \sum_{i\in[n]} \frac{s\|x_s - a^{(i)}\|_2}{(1 + g_s^{(i)})g_s^{(i)}} d_s \leq \int_t^{t'} \frac{5w_s}{t^2 \cdot w_t} d_s \leq \int_t^{t'} \frac{5}{t^2} \cdot \left(\frac{s}{t}\right)^2 ds \\
&= \frac{5}{3t^4}[(t')^2 - (t)^3] = \frac{5}{3t}\left[(1 + \beta)^3 - 1\right] \leq \frac{6\beta}{t} \leq \frac{1}{100t}.
\end{aligned}$$

□

B.2 Line Searching

Here we prove the main results we use on centering, Lemma 10, and line searching Lemma 11. These results are our main tools for computing approximations to the central path. To prove Lemma 11 we also include here two preliminary lemmas, Lemma 22 and Lemma 23, on the structure of $g_{t,y,v}$ defined in (4.1).

Lemma 10. *Given some $y \in \mathbb{R}^d$, $t > 0$ and $0 \leq \epsilon \leq \left(\frac{\mu_t(x)}{8t^2 \cdot w_t(x)}\right)^2$. In $O(nd\log(\frac{n}{\epsilon}))$ time $\texttt{LocalCenter}(y, t, \epsilon)$ computes $x^{(k)}$ such that with high probability in n/ϵ.*

$$f_t(x^{(k)}) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x) \leq \epsilon\left(f_t(y) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x)\right).$$

Proof of Lemma 10. By Lemma 7 we know that $\frac{1}{4}\mathbf{Q} \preceq \nabla^2 f_t(y) \preceq 4\mathbf{Q}$ with high probability in n/ϵ. Furthermore for x such that $\|x-y\|_2 \leq \frac{1}{50t}$ Lemma 1 shows that $\frac{1}{2}\nabla^2 f_t(x) \preceq \nabla^2 f_t(y) \preceq 2\nabla^2 f_t(x)$. Combining these we have that $\frac{1}{8}\mathbf{Q} \preceq \nabla^2 f_t(x) \preceq 8\mathbf{Q}$ for all x with $\|x-y\|_2 \leq \frac{1}{50t}$. Therefore, Lemma 30 shows that

$$f_t(x^{(k)}) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x) \leq \left(1 - \frac{1}{64}\right)^k \left(f_t(x^{(0)}) - \min_{\|x-y\|_2 \leq \frac{1}{49t}} f_t(x)\right).$$

The guarantee on $x^{(k)}$ then follows from our choice of k.

For the running time, Lemma 7 showed the cost of `ApproxMinEig` is $O(nd\log(\frac{n}{\epsilon}))$. Using Lemma 32 we see that the cost per iteration is $O(nd)$ and therefore, the total cost of the k iterations is $O(nd\log(\frac{1}{\epsilon}))$. Combining yields the running time. □

Lemma 22. *For $t > 0$, $y \in \mathbb{R}^d$, and unit vector $v \in \mathbb{R}^d$, the function $g_{t,y,v} : \mathbb{R} \to \mathbb{R}$ defined by (4.1) is convex and nt-Lipschitz.*

Proof. Changing variables yields $g_{t,y,v}(\alpha) = \min_{z \in S} f_t(z + \alpha v)$ for $S \stackrel{\text{def}}{=} \{z \in \mathbb{R}^d : \|z-y\|_2 \leq \frac{1}{49t}\}$. Since f_t is convex and S is a convex set, by Lemma 31 we have that $g_{t,y,v}$ is convex.

Next, by Lemma 13, triangle inequality, and the fact that $t\|x-a^{(i)}\|_2 \leq g_t^{(i)}(x)$ we have

$$\|\nabla f_t(x)\|_2 = \left\|\sum_{i\in[n]} \frac{t^2(x-a^{(i)})}{1+g_t^{(i)}(x)}\right\|_2 \leq \sum_{i\in[n]} \frac{t^2\|x-a^{(i)}\|_2}{1+g_t^{(i)}(x)} \leq tn\,. \tag{B.3}$$

Consequently, $f_t(x)$ is nt-Lipschitz, i.e., for all $x, y \in \mathbb{R}^n$ we have $|f_t(x) - f_t(y)| \leq nt\|x-y\|_2$. Now if we consider the set $S_\alpha \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \|x-(y+\alpha v)\|_2 \leq \frac{1}{49t}\}$ then we see that for all $\alpha, \beta \in \mathbb{R}$ there is a bijection from S_α to S_β were every point in the set moves by at most $\|(\alpha-\beta)v\|_2 \leq |\alpha-\beta|$. Consequently, since $g_{t,y,v}(\alpha)$ simply minimizes f_t over S_α we have that $g_{t,y,v}$ is nt-Lipschitz as desired. □

Lemma 23. *Let $\frac{1}{400f(x_*)} \leq t \leq t' \leq (1+\frac{1}{600})t \leq \frac{2n}{\tilde{\epsilon}_*\cdot\tilde{f}_*}$ and let $(u,\lambda) =$ `ApproxMinEig`(y, t, ϵ_v) for $\epsilon_v \leq \frac{1}{8}(\frac{\tilde{\epsilon}_*}{3n})^2$ and $y \in \mathbb{R}^d$ such that $\|y-x_t\|_2 \leq \frac{1}{t}(\frac{\epsilon_v}{36})^{\frac{3}{2}}$. The function $g_{t,'y,v} : \mathbb{R} \to \mathbb{R}$ defined in (4.1) satisfies $g_{t',y,v}(\alpha_*) = \min_\alpha g_{t,y,v}(\alpha) = f_t(x_t)$ for some $\alpha_* \in [-6f(x_*), 6f(x_*)]$.*

Proof. Let $z \in \mathbb{R}^d$ be an arbitrary unit vector and $\beta = \frac{1}{600}$.

If $\mu_t \leq \frac{1}{4}t^2 \cdot w_t$ then by Lemma 8 and our choice of ϵ_v we have that if $z \perp u$ then

$$|\langle z, v_t\rangle|^2 \leq 8\epsilon_v \leq \left(\frac{\tilde{\epsilon}_*}{3n}\right)^2.$$

Now by Lemma 19 and our bound on t' we know that $\max_{\delta\in[t,t']} \frac{t^2\cdot w_\delta}{\mu_\delta} \leq \frac{3n}{\tilde{\epsilon}_*}$ and hence $|\langle z, v_t\rangle| \leq \min_{\delta\in[t,t']} \frac{\mu_\delta}{t^2\cdot w_\delta}$. By Lemma 5, we know that $z^\top(x_{t'} - x_t) \leq \frac{6\beta}{t} \leq \frac{1}{100t}$.

Otherwise, we have $\mu_t \geq \frac{1}{4}t^2 \cdot w_t$ and by Lemma 9 we have $\|x_{t'} - x_t\|_2 \leq \frac{1}{100t}$.

In either case, since $\|y - x_t\|_2 \leq \frac{1}{100t}$, we can reach $x_{t'}$ from y by first moving an Euclidean distance of $\frac{1}{100t}$ to go from y to x_t, then adding some multiple of v, then moving an Euclidean distance of $\frac{1}{100t}$ in a direction perpendicular to v. Since the total movement perpendicular to v is $\frac{1}{100t} + \frac{1}{100t} \leq \frac{1}{49t'}$ we have that $\min_\alpha g_{t',y,v}(\alpha) = f_{t'}(x_{t'})$ as desired.

All that remains is to show that there is a minimizer of $g_{t',y,v}$ in the range $[-6f(x_*), 6f(x_*)]$. However, by Lemma 6 and Lemma 17 we know that

$$\|y - x_{t'}\|_2 \leq \|y - x_t\|_2 + \|x_t - x_*\|_2 + \|x_* - x_{t'}\|_2 \leq \frac{1}{100t} + f(x_*) + f(x_*) \leq 6f(x_*)\,.$$

Consequently, $\alpha_* \in [-6f(x_*), 6f(x_*)]$ as desired. □

Lemma 11. *Let* $\frac{1}{400f(x_*)} \leq t \leq t' \leq (1 + \frac{1}{600})t \leq \frac{2n}{\tilde{\epsilon}_* \cdot \tilde{f}_*}$ *and let* $(\lambda, u) = \mathtt{ApproxMinEig}(y, t, \epsilon_v)$ *for* $\epsilon_v \leq \frac{1}{8}(\frac{\tilde{\epsilon}_*}{3n})^2$ *and* $y \in \mathbb{R}^d$ *such that* $\|y - x_t\|_2 \leq \frac{1}{t}(\frac{\epsilon_v}{36})^{\frac{3}{2}}$. *In* $O(nd \log^2(\frac{n}{\tilde{\epsilon}_* \cdot \epsilon \cdot \epsilon_v}))$ *time and* $O(\log(\frac{n}{\tilde{\epsilon}_* \cdot \epsilon}))$ *calls to the* `LocalCenter`, `LineSearch`(y, t, t', u, ϵ) *outputs* x' *such that* $\|x' - x_{t'}\|_2 \leq \frac{\epsilon}{t'}$ *with high probability in* n/ϵ.

Proof of Lemma 11. By (B.3) we know that $f_{t'}$ is nt' Lipschitz and therefore

$$f_{t'}(y) - \min_{\|x-y\|_2 \leq \frac{1}{49t'}} f_{t'}(x) \leq \frac{nt'}{49t'} = \frac{n}{49}\,.$$

Furthermore, for $\alpha \in [-6f(x_*), 6f(x_*)]$ we know that by Lemma 17

$$\|y + \alpha u - x_{t'}\|_2 \leq \|y - x_t\|_2 + |\alpha| + \|x_t - x_*\|_2 + \|x_{t'} - x_*\|_2 \leq \frac{1}{t} + 8f(x_*)$$

consequently by Lemma 19 we have

$$\frac{(t')^2 \cdot w_{t'}(y + \alpha u)}{\mu_{t'}(y + \alpha u)} \leq \frac{3n}{\tilde{\epsilon}_*} + t'n\|y + \alpha u - x_{t'}\|_2 \leq \frac{3n}{\tilde{\epsilon}_*} + 2n + 8t'nf(x_*) \leq 20n^2\tilde{\epsilon}_*^{-1}\,.$$

Since $\epsilon_O \leq \left(\frac{\tilde{\epsilon}_*}{160n^2}\right)^2$, invoking Lemma 10 yields that $|q(\alpha) - g_{t',y,u}(\alpha)| \leq \frac{n\epsilon_O}{49}$ with high probability in $n/\tilde{\epsilon}_*$ by and each call to `LocalCenter` takes $O(nd \log \frac{n}{\epsilon_O})$ time. Furthermore, by Lemma 22 we have that $g_{t',y,u}$ is a nt'-Lipschitz convex function and by Lemma 23 we have that the minimizer has value $f_{t'}(x_{t'})$ and is achieved in the range $[-6\tilde{f}_*, 6\tilde{f}_*]$. Consequently, combining all these facts and invoking Lemma E.3, i.e. our result on on one dimensional function minimization, we have $f_{t'}(x') - f_{t'}(x_{t'}) \leq \frac{\epsilon_O}{2}$ using only $O(\log(\frac{nt'f(x_*)}{\epsilon_O}))$ calls to `LocalCenter`.

Finally, by Lemma 20 and Lemma 6 we have

$$\frac{n}{2}\left(\frac{\|x' - x_{t'}\|_2}{\frac{3n/t'}{\tilde{\epsilon}_*} + n\|x' - x_{t'}\|_2}\right)^2 \leq f_{t'}(x') - f_{t'}(x_{t'}) \leq \frac{\epsilon_O}{2}.$$

Hence, we have that

$$\|x' - x_{t'}\|_2 \leq \sqrt{\frac{\epsilon_O}{n}}\left(\frac{3n/t'}{\tilde{\epsilon}_*}\right) + \sqrt{n\epsilon_O}\|x' - x_{t'}\|_2.$$

Since $\epsilon_O = \left(\frac{\epsilon\tilde{\epsilon}_*}{160n^2}\right)^2$, we have

$$\|x' - x_{t'}\|_2 \leq \sqrt{\frac{\epsilon_O}{n}}\frac{6n}{\tilde{\epsilon}_* t'} \leq \frac{\epsilon}{t'}.$$

□

Lemma 12. *Let* $\frac{1}{400f(x_*)} \leq t \leq t' \leq (1 + \frac{1}{600})t \leq \frac{2n}{\tilde{\epsilon}_* \cdot \tilde{f}_*}$ *and let* $x \in \mathbb{R}^d$ *satisfy* $\|x - x_t\|_2 \leq \frac{1}{100t}$. *Then, in* $O(nd \log^2(\frac{n}{\epsilon \cdot \tilde{\epsilon}_*}))$ *time,* `LineSearch`(x, t, t, u, ϵ) *outputs* y *such that* $\|y - x_t\|_2 \leq \frac{\epsilon}{t}$ *for any vector* $u \in \mathbb{R}^d$.

Proof of Lemma 12. The proof is strictly easier than the proof of Lemma 11 as $\|x - \alpha^* u - x_t\|_2 \leq \frac{1}{100t}$ is satisfied automatically for $\alpha^* = 0$. Note that this lemma assume less for the initial point. □

B.3 Putting It All Together

Theorem 1. *In $O(nd\log^3(\frac{n}{\epsilon}))$ time, Algorithm 1 outputs an $(1+\epsilon)$-approximate geometric median with constant probability.*

Proof of Theorem 1. By Lemma 18 we know that $x^{(0)}$ is a 2-approximate geometric median and therefore $f(x^{(0)}) = \tilde{f}_* \leq 2 \cdot f(x_*)$. Furthermore, since $\|x^{(0)} - x_{t_1}\|_2 \leq f(x^{(0)})$ by Lemma 17 and $t_1 = \frac{1}{400\tilde{f}_*}$we have $\|x^{(0)} - x_{t_1}\|_2 \leq \frac{1}{400t_1}$. Hence, by Lemma 12, we have $\|x^{(1)} - x_{t_1}\|_2 \leq \frac{\epsilon_c}{t_1}$ with high probability in n/ϵ. Consequently, by Lemma 11 we have that$\|x^{(k)} - x_{t_i}\|_2 \leq \frac{\epsilon_c}{t_i}$ for all i with high probability in n/ϵ.

Now, Lemma 6 shows that

$$f(x_{t_k}) - f(x^*) \leq \frac{2n}{t_k} \leq \frac{2n}{\tilde{t}_*}\left(1+\frac{1}{600}\right) \leq \tilde{\epsilon}_* \cdot \tilde{f}_* \left(1+\frac{1}{600}\right) \leq \frac{2}{3}\left(1+\frac{1}{600}\right)\epsilon \cdot f(x_*)\,.$$

Since $\|x^{(k)} - x_{t_k}\|_2 \leq \frac{\epsilon_c}{t_k} \leq 400 \cdot \tilde{f}_* \cdot \epsilon_c$ we have that $f(x_k) \leq f(x_{t_k}) + 400n \cdot \tilde{f}_* \cdot \epsilon_c$ by triangle inequality. Combining these facts and using that ϵ_c is sufficiently small yields that $f(x^{(k)}) \leq (1+\epsilon) f(x_*)$ as desired.

To bound the running time, Lemma 7 shows `ApproxMinEvec` takes $O(nd\log(\frac{n}{\epsilon}))$ per iteration and Lemma 11 shows `LineSearch` takes $O\left(nd\log^2\left(\frac{n}{\epsilon}\right)\right)$ time per iteration, using that ϵ_v and ϵ_c are $O(\Omega(\epsilon/n))$. Since for $l = \Omega(\log\frac{n}{\epsilon})$ we have that $t_l > \tilde{t}_*$ we have that $k = O(\log\frac{n}{\epsilon})$. $t_{i+1} \leq \frac{1}{400}$. Since there are $O(\log(\frac{n}{\epsilon}))$ iterations taking time $O\left(nd\log^2\left(\frac{n}{\epsilon}\right)\right)$the running time follows. □

C Pseudo Polynomial Time Algorithm

Here we provide a self-contained result on computing a $1+\epsilon$ approximate geometric median in $O(d\epsilon^{-2})$ time. Note that it is impossible to achieve such approximation for the mean, $\min_{x\in\mathbb{R}^d}\sum_{i\in[n]}\|x - a^{(i)}\|_2^2$, because the mean can be changed arbitrarily by changing only 1 point. However, [19] showed that the geometric median is far more stable. In Section C.1, we show how this stability property allows us to get an constant approximate in $O(d)$ time. In Section C.2, we show how to use stochastic subgradient descent to then improve the accuracy.

C.1 A Constant Approximation of Geometric Median

We first prove that the geometric median is stable even if we are allowed to modify up to half of the points. The following lemma is a strengthening of the robustness result in [19].

Lemma 24. *Let x_* be a geometric median of $\{a^{(i)}\}_{i\in[n]}$ and let $S \subseteq [n]$ with $|S| < \frac{n}{2}$. For all x*

$$\|x_* - x\|_2 \leq \left(\frac{2n - 2|S|}{n - 2|S|}\right)\max_{i\notin S}\|a^{(i)} - x\|_2\,.$$

Proof. For notational convenience let $r = \|x_* - x\|_2$ and let $M = \max_{i\notin S}\|a^{(i)} - x\|_2$.

For all $i \notin S$, we have that $\|x - a^{(i)}\|_2 \leq M$, hence, we have

$$\begin{aligned}\|x_* - a^{(i)}\|_2 &\geq r - \|x - a^{(i)}\|_2\\ &\geq r - 2M + \|x - a^{(i)}\|_2\,.\end{aligned}$$

Furthermore, by triangle inequality for all $i \in S$, we have

$$\|x_* - a^{(i)}\|_2 \geq \|x - a^{(i)}\|_2 - r\,.$$

Hence, we have that

$$\sum_{i\in[n]} \|x_* - a^{(i)}\|_2 \geq \sum_{i\in[n]} \|x - a^{(i)}\|_2 + (n - |S|)(r - 2M) - |S|r\,.$$

Since x_* is a minimizer of $\sum_{i\in[n]} \|x_* - a^{(i)}\|_2$, we have that

$$(n - |S|)(r - 2M) - |S|r \leq 0.$$

Hence, we have

$$\|x_* - x\|_2 = r \leq \frac{2n - 2|S|}{n - 2|S|} M.$$

□

Now, we use Lemma 24 to show that the algorithm `CrudeApproximate` outputs a constant approximation of the geometric median with high probability.

Algorithm 6: $\texttt{CrudeApproximate}_K$

```
Input: $a^{(1)}, a^{(2)}, \cdots, a^{(n)} \in \mathbb{R}^d$.
Sample two independent random subset of $[n]$ of size $K$. Call them $S_1$ and $S_2$.
Let $i^* \in \arg\min_{i\in S_2} \alpha_i$ where $\alpha_i$ is the 65 percentile of the numbers $\{\|a^{(i)} - a^{(j)}\|_2\}_{j\in S_1}$.
Output: Output $a^{(i^*)}$ and $\alpha_{i^*}$.
```

Lemma 25. *Let x_* be a geometric median of $\{a^{(i)}\}_{i\in[n]}$ and $(\widetilde{x}, \lambda)$ be the output of* $\texttt{CrudeApproximate}_K$*. We define $d_T^k(x)$ be the k-percentile of $\left\{\|x - a^{(i)}\|\right\}_{i\in T}$. Then, we have that $\|x_* - \widetilde{x}\|_2 \leq 6d_{[n]}^{60}(\widetilde{x})$. Furthermore, with probability $1 - e^{-\Theta(K)}$, we have*

$$d_{[n]}^{60}(\widetilde{x}) \leq \lambda = d_{S_1}^{65}(\widetilde{x}) \leq 2d_{[n]}^{70}(x_*).$$

Proof. Lemma 24 shows that for all x and $T \subseteq [n]$ with $|T| \leq \frac{n}{2}$

$$\|x_* - x\|_2 \leq \left(\frac{2n - 2|T|}{n - 2|T|}\right) \max_{i\notin T} \|a^{(i)} - x\|_2\,.$$

Picking T to be the indices of largest 40% of $\|a^{(i)} - \widetilde{x}\|_2$, we have

$$\|x_* - \widetilde{x}\|_2 \leq \left(\frac{2n - 0.8n}{n - 0.8n}\right) d_{[n]}^{60}(\widetilde{x}) = 6d_{[n]}^{60}(\widetilde{x}). \tag{C.1}$$

For any point x, we have that $d_{[n]}^{60}(x) \leq d_{S_1}^{65}(x)$ with probability $1 - e^{-\Theta(K)}$ because S_1 is a random subset of $[n]$ with size K. Taking union bound over elements on S_2, with probability $1 - Ke^{-\Theta(K)} = 1 - e^{-\Theta(K)}$, for all points $x \in S_2$

$$d_{[n]}^{60}(x) \leq d_{S_1}^{65}(x). \tag{C.2}$$

yielding that $d^{60}_{[n]}(\widetilde{x}) \leq \lambda$.

Next, for any $i \in S_2$, we have

$$\|a^{(i)} - a^{(j)}\|_2 \leq \|a^{(i)} - x_*\|_2 + \|x_* - a^{(j)}\|_2.$$

and hence

$$d^{70}_{[n]}(a^{(i)}) \leq \|a^{(i)} - x_*\|_2 + d^{70}_{[n]}(x_*).$$

Again, since S_1 is a random subset of $[n]$ with size K, we have that $d^{65}_{S_1}(a^{(i)}) \leq d^{70}_{[n]}(a^{(i)})$ with probability $1 - Ke^{-\Theta(K)} = 1 - e^{-\Theta(K)}$. Therefore,

$$d^{65}_{S_1}(a^{(i)}) \leq \|a^{(i)} - x_*\|_2 + d^{70}_{[n]}(x_*).$$

Since S_2 is an independent random subset, with probability $1 - e^{-\Theta(K)}$, there is $i \in S_2$ such that $\|a^{(i)} - x_*\|_2 \leq d^{70}_{[n]}(x_*)$. In this case, we have

$$d^{65}_{S_1}(a^{(i)}) \leq 2d^{70}_{[n]}(x_*).$$

Since i^* minimize $d^{65}_{S_1}(a^{(i)})$ over all $i \in S_2$, we have that

$$\lambda \overset{\text{def}}{=} d^{65}_{S_1}(\widetilde{x}) \overset{\text{def}}{=} d^{65}_{S_1}(a^{(i^*)}) \leq d^{65}_{S_1}(a^{(i)}) \leq 2d^{70}_{[n]}(x_*)\,.$$

□

C.2 A $1+\epsilon$ Approximation of Geometric Median

Here we show how to improve the constant approximation in the previous section to a $1+\epsilon$ approximation. Our algorithm is essentially stochastic subgradient where we use the information from the previous section to bound the domain in which we need to search for a geometric median.

Algorithm 7: `ApproximateMedian`$(a^{(1)}, a^{(2)}, \cdots, a^{(n)}, \epsilon)$

```
Input: $a^{(1)}, a^{(2)}, \cdots, a^{(n)} \in \mathbb{R}^d$.
Let $T = (60/\epsilon)^2$ and let $\eta = \frac{6\lambda}{n}\sqrt{\frac{2}{T}}$ .
Let $(x^{(1)}, \lambda) = \texttt{CrudeApproximate}_{\sqrt{T}}(a^{(1)}, a^{(2)}, \cdots, a^{(n)})$ .
for $k \leftarrow 1, 2, \cdots, T$ do
    Sample $i_k$ from $[n]$ and let
        $g^{(k)} = \begin{cases} n(x^{(k)} - a^{(i_k)})/\|x^{(k)} - a^{(i_k)}\|_2 & \text{if } x^{(i)} \neq a^{(i_k)} \\ 0 & \text{otherwise} \end{cases}$
    Let $x^{(k+1)} = \arg\min_{\|x - x^{(1)}\|_2 \leq 6\lambda} \eta \left\langle g^{(k)}, x - x^{(k)} \right\rangle + \frac{1}{2}\|x - x^{(k)}\|_2^2$.
end
Output: Output $\frac{1}{T}\sum_{i=1}^{T} x^{(k)}$.
```

Theorem 2. *Let x be the output of* `ApproximateMedian`. *With probability $1 - e^{-\Theta(1/\epsilon)}$, we have*

$$\mathbb{E}f(x) \leq (1+\epsilon) \min_{x \in \mathbb{R}^d} f(x).$$

Furthermore, the algorithm takes $O(d/\epsilon^2)$ time.

Proof. After computing $x^{(1)}$ and λ the remainder of our algorithm is the stocastic subgradient descent method applied to $f(x)$. It is routine to check that $\mathbb{E}_{i^{(k)}}g^{(k)}$ is a subgradient of f at $x^{(k)}$. Furthermore, since the diameter of the domain, $\left\{x : \|x - x^{(1)}\|_2 \leq 6\lambda\right\}$, is clearly λ and the norm of sampled gradient, $g^{(k)}$, is at most n, we have that

$$\mathbb{E}f\left(\frac{1}{T}\sum_{i=1}^{T}x^{(k)}\right) - \min_{\|x-x^{(1)}\|_2\leq 6\lambda} f(x) \leq 6n\lambda\sqrt{\frac{2}{T}}$$

(see [5, Thm 6.1]). Lemma 25 shows that $\|x^* - x^{(1)}\|_2 \leq 6\lambda$ and $\lambda \leq 2d_{[n]}^{70}(x^*)$ with probability $1 - \sqrt{T}e^{-\Theta(\sqrt{T})}$. In this case, we have

$$\mathbb{E}f\left(\frac{1}{T}\sum_{i=1}^{T}x^{(k)}\right) - f(x^*) \quad \leq \quad \frac{12\sqrt{2}nd_{[n]}^{70}(x_*)}{\sqrt{T}}.$$

Since $d_{[n]}^{70}(x^*) \leq \frac{1}{0.3n}f(x^*)$, we have

$$\mathbb{E}f\left(\frac{1}{T}\sum_{i=1}^{T}x^{(k)}\right) \quad \leq \quad \left(1 + \frac{60}{\sqrt{T}}\right)f(x_*) \leq (1+\epsilon)\,f(x_*)\,.$$

□

D Derivation of Penalty Function

Here we derive our penalized objective function. Consider the following optimization problem:

$$\min_{x\in\mathbb{R}^d,\alpha\geq 0\in\mathbb{R}^n} f_t(x,\alpha) \quad \text{where} \quad t\cdot 1^T\alpha + \sum_{i\in[n]} -\ln\left(\alpha_i^2 - \|x-a^{(i)}\|_2^2\right) \quad .$$

Since $p_i(\alpha,x) \stackrel{\text{def}}{=} -\ln\left(\alpha_i^2 - \|x-a^{(i)}\|_2^2\right)$ is a barrier function for the set $\alpha_i^2 \geq \|x-a^{(i)}\|_2^2$, i.e. as $\alpha_i \to \|x-a^{(i)}\|_2$ we have $p_i(\alpha,x) \to \infty$, we see that as we minimize $f_t(x,\alpha)$ for increasing values of t the x values converge to a solution to the geometric median problem. Our penalized objective function, $f_t(x)$, is obtain simply by minimizing the α_i in the above formula and dropping terms that do not affect the minimizing x. In the remainder of this section we show this formally.

Fix some $x \in \mathbb{R}^d$ and $t > 0$. Note that for all $j \in [n]$ we have

$$\frac{\partial}{\partial\alpha_j}f_t(x,\alpha) = t - \left(\frac{1}{\alpha_j^2 - \|x-a^{(i)}\|_2^2}\right)2\alpha_j\,.$$

Since $f(x,\alpha)$ is convex in α, the minimum α_j^* must satisfy

$$t\left(\left(\alpha_j^*\right)^2 - \|x-a^{(i)}\|_2^2\right) - 2\alpha_j^* = 0 \quad . \tag{D.1}$$

Solving for such α_j^* under the restriction $\alpha_j^* \geq 0$ we obtain

$$\alpha_j^* = \frac{2+\sqrt{4+4t^2\|x-a^{(i)}\|_2^2}}{2t} = \frac{1}{t}\left[1+\sqrt{1+t^2\|x-a^{(i)}\|_2^2}\right] \quad . \tag{D.2}$$

Using (D.1) and (D.2) we have that

$$\min_{\alpha\geq 0\in\mathbb{R}^n} f_t(x,\alpha) = \sum_{i\in[n]}\left[1+\sqrt{1+t^2\|x-a^{(i)}\|_2^2} - \ln\left[\frac{2}{t^2}\left(1+\sqrt{1+t^2\|x-a^{(i)}\|_2^2}\right)\right]\right] \ .$$

If we drop the terms that do not affect the minimizing x we obtain our penalty function f_t :

$$f_t(x) = \sum_{i\in[n]}\left[\sqrt{1+t^2\|x-a^{(i)}\|_2^2} - \ln\left(1+\sqrt{1+t^2\|x-a^{(i)}\|_2^2}\right)\right] \ .$$

E Technical Facts

Here we provide various technical lemmas we use through the paper.

E.1 Linear Algebra

First we provide the following lemma that shows that any matrix obtained as a non-negative linear combination of the identity minus a rank 1 matrix less than the identity results in a matrix that is well approximated spectrally by the identity minus a rank 1 matrix. We use this lemma to characterize the Hessian of our penalized objective function and thereby imply that it is possible to apply the inverse of the Hessian to a vector with high precision.

Lemma 26. *Let $\mathbf{A} = \sum_i \left(\alpha_i\mathbf{I} - \beta_i a_i a_i^\top\right) \in \mathbb{R}^{d\times d}$ where the a_i are unit vectors and $0 \leq \beta_i \leq \alpha_i$ for all i. Let v denote a unit vector that is the maximum eigenvector of $\sum_i \beta_i a_i a_i^\top$ and let λ denote the corresponding eigenvalue. Then,*

$$\frac{1}{2}\left(\sum_i \alpha_i\mathbf{I} - \lambda vv^\top\right) \preceq \mathbf{A} \preceq \sum_i \alpha_i\mathbf{I} - \lambda vv^\top \,.$$

Proof. Let $\alpha \overset{\text{def}}{=} \sum_i \alpha_i$. Since clearly $v^\top\mathbf{A}v = v^\top\left(\sum_i \alpha_i\mathbf{I} - \lambda vv^\top\right)v$ it suffices to show that for $w \perp v$ it is the case that $\frac{1}{2}\alpha\|w\|_2^2 \preceq w^\top\mathbf{A}w \preceq \alpha\|w\|_2^2$ or equivalently, that $\lambda_i(\mathbf{A}) \in [\frac{1}{2}\alpha, \alpha]$ for $i \neq d$. However we know that $\sum_{i\in[d]}\lambda_i(\mathbf{A}) = \mathrm{tr}(\mathbf{A}) = d\alpha - \sum_i \beta_i \geq (d-1)\alpha$ and $\lambda_i(\mathbf{A}) \leq \alpha$ for all $i \in [d]$. Consequently, since $\lambda_d(\mathbf{A}) \in [0, \lambda_{d-1}(\mathbf{A})]$ we have

$$2\cdot\lambda_{d-1}(\mathbf{A}) \geq (d-1)\alpha - \sum_{i=1}^{d-2}\lambda_i(\mathbf{A}) \geq (d-1)\alpha - (d-2)\alpha = \alpha \,.$$

Consequently, $\lambda_{d-1}(\mathbf{A}) \in [\frac{\alpha}{2}, \alpha]$ and the result holds by the monotonicity of λ_i. □

Next we bound the spectral difference between the outer product of two unit vectors by their inner product. We use this lemma to bound the amount of precision required in our eigenvector computations.

Lemma 27. *For unit vectors u_1 and u_2 we have*

$$\|u_1u_1^\top - u_2u_2^\top\|_2^2 = 1 - (u_1^\top u_2)^2 \tag{E.1}$$

Consequently if $\left(u_1^\top u_2\right)^2 \geq 1 - \epsilon$ for $\epsilon \leq 1$ we have that

$$-\sqrt{\epsilon}\mathbf{I} \preceq u_1u_1^\top - u_2u_2^\top \preceq \sqrt{\epsilon}\mathbf{I}$$

Proof. Note that $u_1 u_1^\top - u_2 u_2^\top$ is a symmetric matrix and all eigenvectors are either orthogonal to both u_1 and u_2 (with eigenvalue 0) or are of the form $v = \alpha u_1 + \beta u_2$ where α and β are real numbers that are not both 0. Thus, if v is an eigenvector of non-zero eigenvalue λ it must be that

$$\begin{aligned} \lambda \left(\alpha u_1 + \beta u_2\right) &= \left(u_1 u_1^\top - u_2 u_2^\top\right)\left(\alpha u_1 + \beta u_2\right) \\ &= (\alpha + \beta(u_1^\top u_2))u_1 - (\alpha(u_1^\top u_2) + \beta)u_2 \end{aligned}$$

or equivalently

$$\begin{pmatrix} (1-\lambda) & u_1^\top u_2 \\ -(u_1^\top u_2) & -(1+\lambda) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

By computing the determinant we see this has a solution only when

$$-(1-\lambda^2) + (u_1^\top u_2)^2 = 0$$

Solving for λ then yields (E.1) and completes the proof. □

Next we show how the top eigenvectors of two spectrally similar matrices are related. We use this to bound the amount of spectral approximation we need to obtain accurate eigenvector approximations.

Lemma 28. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be symmetric PSD matrices such that* $(1-\epsilon)\mathbf{A} \preceq \mathbf{B} \preceq (1+\epsilon)\mathbf{A}$. *Then if* $g \overset{\text{def}}{=} \frac{\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})}{\lambda_1(\mathbf{A})}$ *satisfies* $g > 0$ *we have* $[v_1(\mathbf{A})^\top v_1(\mathbf{B})]^2 \geq 1 - 2(\epsilon/g)$.

Proof. Without loss of generality $v_1(\mathbf{B}) = \alpha v_1(\mathbf{A}) + \beta v$ for some unit vector $v \perp v_1(\mathbf{A})$ and $\alpha, \beta \in \mathbb{R}$ such that $\alpha^2 + \beta^2 = 1$. Now we know that

$$v_1(\mathbf{B})^\top \mathbf{B} v_1(\mathbf{B}) \leq (1+\epsilon) v_1(\mathbf{B})^\top \mathbf{A} v_1(\mathbf{B}) \leq (1+\epsilon)\left[\alpha^2 \lambda_1(\mathbf{A}) + \beta^2 \lambda_2(\mathbf{A})\right]$$

Furthermore, by the optimality of $v_1(\mathbf{B})$ we have that

$$v_1(\mathbf{B})^\top \mathbf{B} v_1(\mathbf{B}) \geq (1-\epsilon) v_1(\mathbf{A})^\top \mathbf{A} v_1(\mathbf{A}) \geq (1-\epsilon)\lambda_1(\mathbf{A}).$$

Now since $\beta^2 = 1 - \alpha^2$ combining these inequalities yields

$$(1-\epsilon)\lambda_1(\mathbf{A}) \leq (1+\epsilon)\alpha^2\left(\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})\right) + (1+\epsilon)\lambda_2(\mathbf{A}).$$

Rearranging terms, using the definition of g, and that $g \in (0,1]$ and $\epsilon \geq 0$ yields

$$\alpha^2 \geq \frac{\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A}) - \epsilon(\lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A}))}{(1+\epsilon)(\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A}))} = 1 - \frac{2\epsilon\lambda_1(\mathbf{A})}{(1+\epsilon)(\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A}))} \geq 1 - 2(\epsilon/g).$$

□

Here we prove a an approximate transitivity lemma for inner products of vectors. We use this to bound the accuracy need for certain eigenvector computations.

Lemma 29. *Suppose that we have vectors* $v_1, v_2, v_3 \in \mathbb{R}^n$ *such that* $\langle v_1, v_2 \rangle^2 \geq 1 - \epsilon$ *and* $\langle v_2, v_3 \rangle^2 \geq 1 - \epsilon$ *for* $0 < \epsilon \leq \frac{1}{2}$ *then* $\langle v_1, v_3 \rangle^2 \geq 1 - 4\epsilon$.

Proof. Without loss of generality, we can write $v_1 = \alpha_1 v_2 + \beta_1 w_1$ for $\alpha_1^2 + \beta_1^2 = 1$ and unit vector $w_1 \perp v_2$. Similarly we can write $v_3 = \alpha_3 v_2 + \beta_3 w_3$ for $\alpha_3^2 + \beta_3^2 = 1$ and unit vector $w_3 \perp v_2$. Now, by the inner products we know that $\alpha_1^2 \geq 1 - \epsilon$ and $\alpha_3^2 \geq 1 - \epsilon$ and therefore $|\beta_1| \leq \sqrt{\epsilon}$ and $|\beta_3| \leq \sqrt{\epsilon}$. Consequently, since $\epsilon \leq \frac{1}{2}$, $|\beta_1 \beta_3| \leq \epsilon \leq 1 - \epsilon \leq |\alpha_1 \alpha_3|$, and we have

$$\begin{aligned} \langle v_1, v_3 \rangle^2 &\geq \langle \alpha_1 v_2 + \beta_1 w_1, \alpha_3 v_2 + \beta_3 w_3 \rangle^2 \geq (|\alpha_1 \alpha_3| - |\beta_1 \beta_3|)^2 \\ &\geq (1 - \epsilon - \epsilon)^2 = (1 - 2\epsilon)^2 \geq 1 - 4\epsilon. \end{aligned}$$

□

E.2 Convex Optimization

First we provide a single general lemma about about first order methods for convex optimization. We use this lemma for multiple purposes including bounding errors and quickly compute approximations to the central path.

Lemma 30 ([21])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, let $B \subseteq \mathbb{R}$ be a convex set, and let x_* be a point that achieves the minimum value of f restricted to B. Further suppose that for a symmetric positive definite matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$ we have that $\mu \mathbf{H} \preceq \nabla^2 f(y) \preceq L\mathbf{H}$ for all $y \in B$.Then for all $x \in B$ we have*

$$\frac{\mu}{2}\|x - x_*\|_{\mathbf{H}}^2 \leq f(x) - f(x_*) \leq \frac{L}{2}\|x - x_*\|_{\mathbf{H}}^2$$

and

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

Furthermore, if

$$x^{(1)} = \arg\min_{x \in B}\left[f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle + \frac{L}{2}\|x^{(0)} - x\|_{\mathbf{H}}^2\right]$$

then

$$f(x^{(1)}) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)\left(f(x^{(0)}) - f(x_*)\right) . \tag{E.2}$$

Next we provide a short technical lemma about the convexity of functions that arises naturally in our line searching procedure.

Lemma 31. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function and and let $g(\alpha) \stackrel{\text{def}}{=} \min_{x \in S} f(x + \alpha d)$ for any convex set S and $d \in \mathbb{R}^n$. Then g is convex.*

Proof. Let $\alpha, \beta \in \mathbb{R}$ and define $x_\alpha = \arg\min_{x \in S} f(x + \alpha d)$ and $x_\beta = \arg\min_{x \in S} f(x + \beta d)$. For any $t \in [0, 1]$ we have

$$\begin{aligned} g\left(t\alpha + (1-t)\beta\right) &= \min_{x \in S} f\left(x + (t\alpha + (1-t)\beta\right) \\ &\leq f(tx_\alpha + (1-t)x_\beta + (t\alpha + (1-t)\beta)d) && \text{(Convexity of } S\text{)} \\ &\leq t \cdot f(x_\alpha + \alpha d) + (1-t) \cdot f(x_\beta + \beta \cdot d) && \text{(Convexity of } f\text{)} \\ &= t \cdot g(\alpha) + (1-t) \cdot g(\beta) \end{aligned}$$

□

Lemma 32. *For any vectors $y, z, v \in \mathbb{R}^d$ and scalar α, we can compute $\arg\min_{\|x-y\|_2^2 \leq \alpha} \|x - z\|_{\mathbf{I} - vv^\top}^2$ exactly in time $O(d)$.*

Proof. Let x^* be the solution of this problem. If $\|x^* - y\|_2^2 < \alpha$, then $x^* = z$. Otherwise, there is $\lambda > 0$ such that x^* is the minimizer of

$$\min_{x \in \mathbb{R}^d} \|x - z\|_{\mathbf{I} - vv^\top}^2 + \lambda\|x - y\|_2^2 .$$

Let $\mathbf{Q} = \mathbf{I} - vv^\top$. Then, the optimality condition of the above equation shows that

$$\mathbf{Q}(x^* - z) + \lambda(x^* - y) = 0 .$$

Therefore,

$$x^* = (\mathbf{Q} + \lambda\mathbf{I})^{-1}(\mathbf{Q}z + \lambda y)\,. \tag{E.3}$$

Hence,

$$\alpha = \|x^* - y\|_2^2 = (z-y)^\top \mathbf{Q}(\mathbf{Q}+\lambda\mathbf{I})^{-2}\mathbf{Q}(z-y).$$

Let $\eta = 1 + \lambda$, then we have $(\mathbf{Q} + \lambda\mathbf{I}) = \eta\mathbf{I} - vv^\top$and hence Sherman–Morrison formula shows that

$$(\mathbf{Q}+\lambda\mathbf{I})^{-1} = \eta^{-1}\mathbf{I} + \frac{\eta^{-2}vv^\top}{1 - \|v\|^2\eta^{-1}} = \eta^{-1}\left(\mathbf{I} + \frac{vv^\top}{\eta - \|v\|^2}\right).$$

Hence, we have

$$(\mathbf{Q}+\lambda\mathbf{I})^{-2} = \eta^{-2}\left(\mathbf{I} + \frac{2vv^\top}{\eta - \|v\|^2} + \frac{vv^\top\|v\|^2}{(\eta-\|v\|^2)^2}\right) = \eta^{-2}\left(\mathbf{I} + \frac{2\eta - \|v\|^2}{(\eta-\|v\|^2)^2}vv^\top\right).$$

Let $c_1 = \|\mathbf{Q}(z-y)\|_2^2$ and $c_2 = \left(v^\top\mathbf{Q}(z-y)\right)^2$, then we have

$$\alpha \;=\; \eta^{-2}\left(c_1 + \frac{2\eta - \|v\|^2}{(\eta-\|v\|^2)^2}c_2\right).$$

Hence, we have

$$\alpha\eta^2\left(\eta - \|v\|^2\right)^2 = c_1\left(\eta - \|v\|^2\right)^2 + c_2\left(2\eta - \|v\|^2\right).$$

Note that this is a polynomial of degree 4 in η and all coefficients can be computed in $O(d)$ time. Solving this by explicit formula, one can test all 4 possible η's into the formula (E.3) of x. Together with trivial case $x^* = z$, we simply need to check among 5 cases to check which is the solution. □

E.3 Noisy One Dimensional Convex Optimization

Here we show how to minimize a one dimensional convex function giving a noisy oracle for evaluating the function. While this could possibly be done using general results on convex optimization with a membership oracle, the proof in one dimension is much simpler and we provide it here for completeness.

Lemma 33. *Let $f : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz convex function defined on the $[\ell, u]$ interval and let $g : \mathbb{R} \to \mathbb{R}$ be an oracle such that $|g(y) - f(y)| \leq \epsilon$ for all y. In $O(\log(\frac{L(u-\ell)}{\epsilon}))$ time and with $O(\log(\frac{L(u-\ell)}{\epsilon}))$ calls to g, the algorithm* `OneDimMinimizer`$(\ell, u, \epsilon, g, L)$ *outputs a point x such that*

$$f(x) - \min_{y \in [\ell, u]} f(y) \leq 4\epsilon.$$

Proof. First, note that for any $y, y' \in \mathbb{R}$ if $f(y) < f(y') - 2\epsilon$ then $g(y) < g(y')$. This directly follows from our assumption on g. Second, note that the output of the algorithm, x, is simply the point queried by the algorithm (i.e. ℓ and the z_ℓ^i and z_u^i) with the smallest value of g. Combining these facts implies that $f(x)$ is within 2ϵ of the minimum value of f among the points queried. It thus suffices to show that the algorithm queries some point within 2ϵ of optimal.

To do this, we break into two cases. First, consider the case where the intervals $[y_\ell^{(i)}, y_u^{(i)}]$ all contain a minimizer of f. In this case, the final interval contains an optimum, and is of size at most $\frac{\epsilon}{L}$. Thus, by the Lipschitz property, all points in the interval are within $\epsilon \leq 2\epsilon$ of optimal, and at least one endpoint of the interval must have been queried by the algorithm.

```
Algorithm 8: OneDimMinimizer$(\ell, u, \epsilon, g, L)$
  Input: Interval $[\ell, u] \subseteq \mathbb{R}$ and target additive error $\epsilon \in \mathbb{R}$
  Input: noisy additive evaluation oracle $g : \mathbb{R} \to \mathbb{R}$ and Lipschitz bound $L > 0$
  Let $x^{(0)} = \ell$, $y_\ell^{(0)} = \ell$, $y_u^{(0)} = u$
  for $i = 1, ..., \left\lceil \log_{3/2}(\frac{L(u-\ell)}{\epsilon}) \right\rceil$ do
      Let $z_\ell^{(i)} = \frac{2y_\ell^{(i-1)} + y_u^{(i-1)}}{3}$ and $z_u^{(i)} = \frac{y_\ell^{(i-1)} + 2y_u^{(i-1)}}{3}$
      if $g(z_\ell^{(i)}) \leq g(z_u^{(i)})$ then
          Let $(y_\ell^{(i)}, y_u^{(i)}) = (y_\ell^{(i-1)}, z_u^{(i)})$.
          If $g(z_\ell^{(i)}) \leq g(x^{(i-1)})$ update $x^{(i)} = z_\ell^{(i)}$..
      else if $g(z_\ell^{(i)}) > g(z_u^{(i)})$ then
          Let $(y_\ell^{(i)}, y_u^{(i)}) = (z_\ell^{(i)}, y_u^{(i-1)})$.
          If $g(z_u^{(i)}) \leq g(x^{(i-1)})$ update $x^{(i)} = z_u^{(i)}$.
      end
  end
  Output: $x^{(\mathrm{last})}$
```

For the other case, consider the last i for which this interval does contain an optimum of f. This means that $g(z_\ell^{(i)}) \leq g(z_u^{(i)})$ while a minimizer x^* is to the right of $z_u^{(i)}$, or the symmetric case with a minimizer is to the left of $z_\ell^{(i)}$. Without loss of generality, we assume the former. We then have $z_\ell^{(i)} \leq z_u^{(i)} \leq x^*$ and $x^* - z_u^{(i)} \leq z_u^{(i)} - z_\ell^{(i)}$. Consequently $z_u^{(i)} = \alpha z_l^{(i)} + (1-\alpha)x^*$ where $\alpha \in [0, \frac{1}{2}]$ and the convexity of f implies $f(z_u^{(i)}) \leq \frac{1}{2} f(z_l^{(i)}) + \frac{1}{2} f(x^*)$ or equivalently $f(z_u^{(i)}) - f(x^*) \leq f(z_\ell^{(i)}) - f(z_u^{(i)})$. But $f(z_\ell^{(i)}) - f(z_u^{(i)}) \leq 2\epsilon$ since $g(z_\ell^{(i)}) \leq g(z_u^{(i)})$. Thus, $f(z_u^{(i)}) - f(x^*) \leq 2\epsilon$, and $z_u^{(i)}$ is queried by the algorithm, as desired. □

F Weighted Geometric Median

In this section, we show how to extend our results to the *weighted geometric median* problem, also known as the Weber problem: given a set of n points in d dimensions, $a^{(1)}, \ldots, a^{(n)} \in \mathbb{R}^d$, with corresponding weights $w^{(1)}, \ldots, w^{(n)} \in \mathbb{R}_{>0}$, find a point $x_* \in \mathbb{R}^d$ that minimizes the weighted sum of Euclidean distances to them:

$$x_* \in \arg\min_{x \in \mathbb{R}^d} f(x) \quad \text{where} \quad f(x) \stackrel{\text{def}}{=} \sum_{i \in [n]} w^{(i)} \|x - a^{(i)}\|_2.$$

As in the unweighted problem, our goal is to compute $(1+\epsilon)$-approximate solution, i.e. $x \in \mathbb{R}^d$ with $f(x) \leq (1+\epsilon) f(x_*)$.

First, we show that it suffices to consider the case where the weights are integers with bounded sum (Lemma 34). Then, we show that such an instance of the weighted geometric median problem can be solved using the algorithms developed for the unweighted problem.

Lemma 34. *Given points $a^{(1)}, a^{(2)}, \ldots, a^{(n)} \in \mathbb{R}^d$, non-negative weights $w^{(1)}, w^{(2)}, \ldots, w^{(n)} \in \mathbb{R}_{>0}$, and $\epsilon \in (0,1)$, we can compute in linear time weights $w_1^{(1)}, w_1^{(2)}, \ldots, w_1^{(n)}$ such that:*

- *Any* $(1+\epsilon/5)$*-approximate weighted geometric median of* $a^{(1)}, \dots, a^{(n)}$ *with the weights* $w_1^{(1)}, \dots, w_1^{(n)}$ *is also a* $(1+\epsilon)$*-approximate weighted geometric median of* $a^{(1)}, \dots, a^{(n)}$ *with the weights* $w^{(1)}, \dots, w^{(n)}$*, and*
- $w_1^{(1)}, \dots, w_1^{(n)}$ *are nonnegative integers and* $\sum_{i \in [n]} w_1^{(i)} \le 5n\epsilon^{-1}$.

Proof. Let

$$f(x) = \sum_{i \in [n]} w^{(i)} \|a^{(i)} - x\|$$

and $W = \sum_{i \in [n]} w^{(i)}$. Furthermore, let $\epsilon' = \epsilon/5$ and for each $i \in [n]$, define $w_0^{(i)} = \frac{n}{\epsilon' W} w^{(i)}$, $w_1^{(i)} = \left\lfloor w_0^{(i)} \right\rfloor$ and $w_2^{(i)} = w_0^{(i)} - w_1^{(i)}$. We also define $f_0, f_1, f_2, W_0, W_1, W_2$ analogously to f and W.

Now, assume $f_1(x) \le (1+\epsilon')f_1(x_*)$, where x_* is the minimizer of f and f_0. Then:

$$f_0(x) = f_1(x) + f_2(x) \le f_1(x) + f_2(x_*) + W_2\|x - x_*\|_2$$

and

$$\begin{aligned} W_2\|x - x_*\|_2 &= \frac{W_2}{W_1} \sum_{i \in [n]} w_1^{(i)} \|x - x_*\|_2 \le \frac{W_2}{W_1} \sum_{i \in [n]} w_1^{(i)} \left(\|x - a^{(i)}\|_2 + \|a^{(i)} - x_*\| \right) \\ &\le \frac{W_2}{W_1} \left(f_1(x) + f_1(x_*) \right). \end{aligned}$$

Now, since $W_0 = \frac{n}{\epsilon'}$ and $W_1 \ge W_0 - n$ we have

$$\frac{W_2}{W_1} = \frac{W_0 - W_1}{W_1} = \frac{W_0}{W_1} - 1 \le \frac{W_0}{W_0 - n} - \frac{W_0 - n}{W_0 - n} = \frac{n}{\frac{n}{\epsilon'} - n} = \frac{\epsilon'}{1 - \epsilon'}.$$

Combining these yields that

$$\begin{aligned} f_0(x) &\le f_1(x) + f_2(x_*) + \frac{\epsilon'}{1-\epsilon'} \left(f_1(x) + f_1(x_*) \right) \\ &\le \left(1 + \frac{\epsilon'}{1-\epsilon'}\right)(1+\epsilon') f_1(x^*) + \frac{\epsilon'}{1-\epsilon'} f_1(x_*) + f_2(x_*) \\ &\le (1 + 5\epsilon') f_0(x_*) = (1+\epsilon) f_0(x_*). \end{aligned}$$

□

We now proceed to show the main result of this section.

Lemma 35. *A* $(1+\epsilon)$*-approximate weighted geometric median of* n *points in* $\mathbb{R}^d$ *can be computed in* $O(nd \log^3 \epsilon^{-1})$ *time.*

Proof. By applying Lemma 34, we can assume that the weights are integer and their sum does not exceed $n\epsilon^{-1}$. Note that computing the weighted geometric median with such weights is equivalent to computing an unweighted geometric median of $O(n\epsilon^{-1})$ points (where each point of the original input is repeated with the appropriate multiplicity). We now show how to simulate the behavior of our unweighted geometric median algorithms on such a set of points without computing it explicitly.

If $\epsilon > n^{-1/2}$, we will apply the algorithm `ApproximateMedian`, achieving a runtime of $O(d\epsilon^{-2}) = O(nd)$. It is only necessary to check that we can implement weighted sampling from our points with $O(n)$ preprocessing and $O(1)$ time per sample. This is achieved by the alias method [15].

Now assume $\epsilon < n^{-1/2}$. We will employ the algorithm `AccurateMedian`. Note that we can implement the subroutines `LineSearch` and `ApproxMinEig` on the implicitly represented multiset of $O(n\epsilon^{-1})$ points. It is enough to observe only n of the points are distinct, and all computations performed by these subroutines are identical for identical points. The total runtime will thus be $O(nd \log^3(n/\epsilon^2)) = O(nd \log^3 \epsilon^{-1})$. □