Highly smooth minimization of non-smooth problems

Brian Bullins BBULLINS@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We establish improved rates for structured *non-smooth* optimization problems by means of near-optimal higher-order accelerated methods. In particular, given access to a standard oracle model that provides a p^{th} order Taylor expansion of a *smoothed* version of the function, we show how to achieve ε-optimality for the *original* problem in $\tilde{O}_p\left(\varepsilon^{-\frac{2p+2}{3p+1}}\right)$ calls to the oracle. Furthermore, when $p = 3$, we provide an efficient implementation of the near-optimal accelerated scheme that achieves an $O(\varepsilon^{-4/5})$ iteration complexity, where each iteration requires $\tilde{O}(1)$ calls to a linear system solver. Thus, we go beyond the previous $O(\varepsilon^{-1})$ barrier in terms of ε dependence, and in the case of ℓ_∞ regression and ℓ_1-SVM, we establish overall improvements for some parameter settings in the moderate-accuracy regime. Our results also lead to improved high-accuracy rates for minimizing a large class of convex quartic polynomials.

Keywords: Non-smooth convex optimization, higher-order acceleration, ℓ_∞ regression

1. Introduction

While the benefit of smoothness for improved convergence guarantees is well understood in the optimization literature, many problems of interest are unfortunately non-smooth, and thus do not inherit these favorable rates. One such example is the classic problem of ℓ_∞ regression:

$$\min_{x\in\mathbb{R}^d} \|\mathbf{A}x - b\|_\infty, \quad \mathbf{A} \in \mathbb{R}^{m\times d},\ b \in \mathbb{R}^m. \tag{1}$$

Although a first-order iteration complexity of $O(1/\varepsilon^2)$ can be obtained when optimizing Lipschitz continuous convex functions, it is known that one can achieve better than the black-box rate for certain structured functions (Nemirovski, 2004; Nesterov, 2005b,a, 2007), such as ℓ_∞ and ℓ_1 regression, as well as bilinear saddle-point problems.

In this work, we go beyond these previous first-order approaches to establish improved higher-order smoothed oracle complexities for several important non-smooth optimization problems, including ℓ_∞ regression. As noted by Ene and Vladu (2019), even achieving a linear dependence in ε^{-1} has required careful handling of accelerated techniques for non-smooth optimization (Nesterov, 2005b; Sherman, 2017; Sidford and Tian, 2018). Thus, we show how to go *beyond* these rates to achieve oracle complexities *sublinear* in ε^{-1}. We further extend these results to the setting of ℓ_1-SVM, again achieving oracle complexities that are sublinear in ε^{-1}. Additionally, under third-order smoothness assumptions (i.e., the $p = 3$ case), we make use of efficient tensor methods Nesterov (2018a) in order to establish overall computational costs in terms of (per-iteration) linear system solves, thus providing results that may be compared with (Christiano et al., 2011; Chin et al., 2013;

Ene and Vladu, 2019), where the ℓ_∞ regression problem has been considered in the context of approximate max flow.

An important observation of this work is that the softmax approximation to the max function, which we denote as $\mathrm{smax}_\mu(\cdot)$ (parameterized by $\mu > 0$), is not only smooth (i.e., its gradient is Lipschitz continuous), but also *higher-order smooth*. In particular, we establish Lipschitz continuity of its p^{th} derivatives with Lipschitz constant $O_p(1/\mu^p)$, where we use $O_p(\cdot)$ to hide additional p-dependent terms. By combining this observation with recent advances in higher-order acceleration (Gasnikov et al., 2018; Jiang et al., 2018; Bubeck et al., 2018b; Bullins, 2018; Gasnikov et al., 2019), we achieve an improved p^{th}-order oracle complexity of $\tilde{O}_p(\varepsilon^{-\frac{2p+2}{3p+1}})$, thus establishing a family of rates that goes beyond the previous $O(1/\varepsilon)$ dependence for $p > 1$ (Nesterov, 2005b; Sherman, 2017; Sidford and Tian, 2018; Ene and Vladu, 2019).

1.1. Our contributions

The main contributions of this work are as follows:

1. We provide improved higher-order oracle complexities for several important *non-smooth* optimization problems, by combining near-optimal higher-order acceleration with the appropriate highly smooth approximations.

2. By leveraging efficient tensor methods for the case when $p = 3$ (Nesterov, 2018a), we go beyond the oracle model to establish overall computational cost for these non-smooth problems that, for certain parameter regimes (see: Appendix A), improves upon previous results.

3. Our efficient tensor methods can further be extended to the high-accuracy regime, whereby we show in Appendix C improved convergence rates for a large class of convex quartic polynomials. By doing so, we arrive at a convergence rate for ℓ_4 regression that improves upon the rate of Bubeck et al. (2018a), and matches that of Adil et al. (2019a) (up to logarithmic factors).

1.2. Overview of approach

We begin by considering the value that softmax provides as an approximation to the (non-smooth) max function. In particular, we go beyond its standard first-order smoothness to instead show how to bound its p^{th}-order derivatives for all orders $p \geq 1$, as a function of p. Ultimately, the higher-order smoothness guarantees combine with near-optimal higher-order accelerated methods (Gasnikov et al., 2018; Jiang et al., 2018; Bubeck et al., 2018b; Bullins, 2018; Gasnikov et al., 2019) to result in the higher-order smoothed oracle complexity of $\tilde{O}_p(\varepsilon^{-\frac{2p+2}{3p+1}})$, for $p \geq 1$.

Once we shift to the specific case for $p = 3$, our approach is primarily based on extending a near-optimal accelerated higher-order optimization procedure (Monteiro and Svaiter, 2013), whereby each iteration of the method requires finding an exact minimizer of a subproblem given by the third-order Taylor expansion, centered around the t^{th} iterate, plus an additional fourth-order regularization term. As our aim is to go beyond the oracle model, we leverage an efficient third-order tensor method (Nesterov, 2018a) which provides a sufficiently accurate solution to the subproblem. We note that the approach presented by Nesterov (2018a) is highly tuned to the fourth-order regularized model, and so extending this type of result beyond fourth-order regularization remains an interesting open question.

After a part of this work first appeared on arXiv (Bullins and Peng, 2019), follow-up work by Carmon et al. (2020) showed how to achieve a rate of $\tilde{O}\left(\|x_0 - x^*\|_{\mathbf{A}^\top\mathbf{A}}^{2/3}\varepsilon^{-2/3}\right)$ for ℓ_∞ regression, by combining Monteiro-Svaiter acceleration with access to an efficiently implementable ball oracle. We believe both works provide further evidence of the value of these acceleration schemes, and we look forward to exploring these promising directions. Due to space constraints, we provide a more extensive overview of related works in the appendix.

1.3. Organization of the paper

Our paper is organized as follows. In Section 2, we establish the necessary definitions and machinery for handling higher-order derivatives, along with the relevant extensions to higher-order notions of smoothness and strong convexity. Then, in Section 3, we present the standard softmax function as a smooth approximation to the max function, whereby we show that its smoothness properties extend to all orders. Combining this result with recent advances in higher-order optimization leads to our main oracle complexity results, Theorems 10 and 11. In Section 4, we focus on the case where $p = 3$, thus allowing us to go beyond the oracle model and arrive at overall computational guarantees in the form of Theorems 12 and 13.

2. Setup

Let u, v denote vectors in $\mathbb{R}^d$. Throughout, we let v_i denote the i-th coordinate of v, and we let $[k] \stackrel{\text{def}}{=} \{1, \ldots, k\}$ for $k \geq 1$. We let $\Delta_m \stackrel{\text{def}}{=} \{x \in \mathbb{R}^m : \sum_i x_i = 1, x_i \geq 0\}$ denote the m-dimensional simplex. We let $\|v\|_p$ denote the standard ℓ_p norm, and we drop the subscript to let $\|\cdot\|$ denote the ℓ_2 norm. Let $\mathbf{B} \in \mathbb{R}^{d\times d}$ be a symmetric positive-definite matrix, i.e., $\mathbf{B} \succ 0$. Then, we may define the matrix-induced norm of v (w.r.t. $\mathbf{B}$) as $\|v\|_{\mathbf{B}} \stackrel{\text{def}}{=} \sqrt{v^\top \mathbf{B} v}$, and we let $\|\mathbf{B}\| \stackrel{\text{def}}{=} \lambda_{\max}(\mathbf{B})$.

We now make formal a higher-order notion of smoothness. Specifically, for $p \geq 1$, we say a p-times differentiable function $f(\cdot)$ is *L_p-smooth (of order p)* w.r.t. $\|\cdot\|_{\mathbf{B}}$ if the p^{th} derivative is L_p-Lipschitz continuous, i.e., for all $x, y \in \mathbb{R}^d$,

$$\|\nabla^p f(y) - \nabla^p f(x)\|_{\mathbf{B}}^* \leq L_p \|y - x\|_{\mathbf{B}}, \tag{2}$$

where we define

$$\|\nabla^p f(y) - \nabla^p f(x)\|_{\mathbf{B}}^* \stackrel{\text{def}}{=} \max_{h:\|h\|_{\mathbf{B}}\leq 1} \left|\nabla^p f(y)[h]^p - \nabla^p f(x)[h]^p\right|,$$

and where

$$\nabla^p f(x)[h]^p \stackrel{\text{def}}{=} \nabla^p f(x)\underbrace{[h, h, \ldots, h]}_{p \text{ times}}.$$

Observe that, for $p = 1$, this recovers the usual notion of smoothness, and so our convention will be to refer to first-order smooth functions as simply smooth. A complementary notion is that of strong convexity, and its higher-order generalization known as uniform convexity (Nesterov, 2008). In particular, $f(\cdot)$ is *σ_p-uniformly convex (of order p)* with respect to $\|\cdot\|_{\mathbf{B}}$ if, for all $x, y \in \mathbb{R}^d$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\sigma_p}{p}\|y - x\|_{\mathbf{B}}^p.$$

Again, we may see that this captures the typical σ_2-strong convexity (w.r.t. $\|\cdot\|_{\mathbf{B}}$) by setting $p = 2$.

Following the conventions of Nesterov (2018a), we define the p^{th}-order Taylor expansion, centered at x, as

$$\Phi_{x,p}(y) \stackrel{\text{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} \nabla^i f(x)[y-x]^i, \; p \geq 1. \tag{3}$$

3. Softmax approximation for non-smooth problems

We recall from (Nesterov, 2005b; Sidford and Tian, 2018) the standard softmax approximation, for $x \in \mathbb{R}^m$:

$$\operatorname{smax}_\mu(x) \stackrel{\text{def}}{=} \mu \log \left(\sum_{i=1}^{m} e^{\frac{x_i}{\mu}} \right). \tag{4}$$

It is straightforward to observe that (4) is $\frac{1}{\mu}$-smooth, and furthermore that it smoothly approximates the max function, i.e., $\max_{j \in [m]} x_j$ (Sidford and Tian, 2018).

Fact 1 *For all $x \in \mathbb{R}^m$,*

$$\max_{j \in [m]} x_j \leq \operatorname{smax}_\mu(x) \leq \mu \log(m) + \max_{j \in [m]} x_j. \tag{5}$$

Note that this approximation can be used to approximate $\|x\|_\infty$, since $\|x\|_\infty = \max\limits_{j \in [m]} |x_j|$, and $|x_j| = \max\{x_j, -x_j\}$. It follows that we may determine a smooth approximation of ℓ_∞ regression, i.e.,

$$\min_{x \in \mathbb{R}^d} \|\tilde{\mathbf{A}} x - \tilde{b}\|_\infty, \quad \tilde{\mathbf{A}} \in \mathbb{R}^{m \times d}, \;\; \tilde{b} \in \mathbb{R}^m, \tag{6}$$

as $\operatorname{smax}_\mu(\mathbf{A}x - b)$, where $\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} \\ -\tilde{\mathbf{A}} \end{pmatrix}$ and $b = \begin{pmatrix} \tilde{b} \\ -\tilde{b} \end{pmatrix}$.

Having now formalized the connection between $\operatorname{smax}_\mu(\cdot)$ and $\|\cdot\|_\infty$, we assume throughout the rest of the section that $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$, as the difference in dimension between $\tilde{\mathbf{A}}$, $\tilde{b}$ and $\mathbf{A}$, b only affects the final convergence by a constant factor. In addition, we will assume that $\mathbf{A}$ is such that $\mathbf{A}^\top \mathbf{A} \succ 0$, and thus we consider the regime where $m \geq d$.

3.1. ℓ_1-regularized SVM

We may also consider the ℓ_1-regularized soft-margin SVM (ℓ_1-SVM) problem, i.e.,

$$f(x) = \lambda \|x\|_1 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - b_i \langle a_i, x \rangle\}, \tag{7}$$

for $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ ($i \in [m]$), and $\lambda > 0$. To simplify the notation, we define

$$\operatorname{SVM}(x) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - x_i\}.$$

Letting $\tilde{q}_i \stackrel{\text{def}}{=} b_i a_i$ and $\tilde{\mathbf{Q}} \stackrel{\text{def}}{=} [\tilde{q}_1 \; \tilde{q}_2 \; \ldots \; \tilde{q}_m]^\top$, we may then rewrite $f(x) = \lambda \|x\|_1 + \operatorname{SVM}(\tilde{\mathbf{Q}} x)$. We now make the following observations concerning softmax-based approximations for $\|\cdot\|_1$ and $\max\{0, \cdot\}$.

Lemma 2 (ℓ_1 approximation) *Let* $\mathrm{sabs}_\mu(c) \overset{def}{=} \mathrm{smax}_\mu([c,-c])$ *for* $c \in \mathbb{R}$*, and let* $\text{soft-}\ell 1_\mu(x) \overset{def}{=} \sum_{i=1}^m \mathrm{sabs}_\mu(x_i)$ *for* $x \in \mathbb{R}^m$*. Then, we have that*

$$\|x\|_1 \leq \text{soft-}\ell 1_\mu(x) \leq \|x\|_1 + \mu m. \tag{8}$$

Lemma 3 (Smooth hinge loss approximation) *Let* $\mathrm{shinge}_\mu(c) \overset{def}{=} \mathrm{smax}_\mu([0,c])$ *for* $c \in \mathbb{R}$*. Then*

$$\max\{0,c\} \leq \mathrm{shinge}_\mu(c) \leq \max\{0,c\} + \mu. \tag{9}$$

This gives us a natural smooth approximation to $\mathrm{SVM}(x)$, namely,

$$\mathrm{softSVM}_\mu(x) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \mathrm{shinge}_\mu(1-x_i). \tag{10}$$

Taken together with these approximations, we arrive at the following lemma, the proof of which follows by combining Lemmas 2 and 3.

Lemma 4 *Let* $f_\mu(x) = \lambda\,\text{soft-}\ell 1_\mu(x) + \mathrm{softSVM}(\tilde{\mathbf{Q}}x)$*, and let* $f(x)$ *be as in* (7)*. Then, for all* $x \in \mathbb{R}^d$*,*

$$f(x) \leq f_\mu(x) \leq f(x) + 2\mu\lambda d. \tag{11}$$

3.2. Softmax calculus and higher-order smoothness

Now that we have established the connection between softmax and some important non-smooth functions, we shift our attention to several desirable properties of $\mathrm{smax}_\mu(\cdot)$. To simplify notation, we let $Z_\mu(x) = \sum_{i=1}^m e^{\frac{x_i}{\mu}}$, and so $\mathrm{smax}_\mu(x) = \mu\log\left(Z_\mu(x)\right)$. Note that we have

$$\nabla\,\mathrm{smax}_\mu(x)_i = \frac{e^{\frac{x_i}{\mu}}}{Z_\mu(x)}, \quad i \in [m]\,. \tag{12}$$

Furthermore, since $\nabla\,\mathrm{smax}_\mu(x) \in \Delta_m$ for all $x \in \mathbb{R}^m$, it follows that, for all $p \geq 1$,

$$\|\nabla\,\mathrm{smax}_\mu(x)\|_p \leq 1. \tag{13}$$

We may also see that

$$\nabla^2\,\mathrm{smax}_\mu(x) = \frac{1}{\mu}\left(\mathrm{diag}(\nabla\,\mathrm{smax}_\mu(x)) - \nabla\,\mathrm{smax}_\mu(x)\nabla\,\mathrm{smax}_\mu(x)^\top\right). \tag{14}$$

As mentioned previously, one of the key observations of this work is that softmax is equipped with favorable higher-order smoothness properties. Thus, the following lemma shows how we may bound its p^{th}-order derivatives, for all $p \geq 1$, and its proof can be found in the appendix.

Theorem 5 *For all* x*,* $h \in \mathbb{R}^d$*,* $p \geq 1$*,*

$$|\nabla^p\,\mathrm{smax}_\mu(x)[h]^p| \leq \frac{\left(\frac{p}{\ln(p+1)}\right)^p (p-1)!\|h\|_2^p}{\mu^{p-1}}. \tag{15}$$

It will also be helpful to note the following standard result on how a bound on the $(p+1)^{th}$ derivative implies Lipschitz-continuity of the p^{th} derivative.

Lemma 6 *Let $f(\cdot)$ be a $(p+1)$-times differentiable function, let $L_p > 0$ and $\mathbf{A}$ be such that $\mathbf{A}^\top \mathbf{A} \succ 0$, and suppose, for all $\zeta, h \in \mathbb{R}^d$,*

$$\left|\nabla^{p+1} f(\zeta)[h]^{p+1}\right| \leq L_p \|\mathbf{A}h\|_2^{p+1}. \tag{16}$$

Then we have that, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla^p f(y) - \nabla^p f(x)\|^*_{\mathbf{A}^\top \mathbf{A}} \leq L_p \|y - x\|_{\mathbf{A}^\top \mathbf{A}}. \tag{17}$$

Having determined these bounds, we now provide smoothness guarantees for the softmax approximation to both ℓ_∞ regression and ℓ_1-SVM.

Theorem 7 *Let $f_\mu(x) = \mathrm{smax}_\mu(\mathbf{A}x - b)$. Then, $f_\mu(x)$ is (order p) $\frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!}{\mu^p}$ - smooth w.r.t. $\|\cdot\|_{\mathbf{A}^\top \mathbf{A}}$.*

Theorem 8 *Let $f_\mu(x) = \lambda\,\text{soft-}\ell 1_\mu(x) + \mathrm{softSVM}_\mu(\tilde{\mathbf{Q}}x)$. Then, $f_\mu(x)$ is (order p) Q-smooth w.r.t. $\|\cdot\|_2$, for $Q = \frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!\left(\lambda d + \|\tilde{\mathbf{Q}}^\top \tilde{\mathbf{Q}}\|^{\frac{p+1}{2}}\right)}{\mu^p}$.*

We now consider recent advances in near-optimal accelerated methods for higher-order smooth convex optimization (Gasnikov et al., 2018; Jiang et al., 2018; Bubeck et al., 2018b; Bullins, 2018; Gasnikov et al., 2019). While we will further explore the details behind the method in Section 4, the overall idea is to combine a carefully tuned acceleration scheme with a regularized p^{th}-order Taylor expansion oracle, whereby the inner step of the acceleration scheme requires minimizing the regularized Taylor model.

Theorem 9 (Bubeck et al. (2018b), Theorem 1.1) *Let $f(\cdot)$ denote a convex function whose p^{th} derivative is L_p-Lipschitz, and let x^* denote a minimizer of $f(\cdot)$. Then, the* Accelerated Taylor Descent (ATD) *method (Bubeck et al. (2018b), Algorithm 1) satisfies, with $c_p = 2^{p-1}(p+1)^{\frac{3p+1}{2}}/(p-1)!$,*

$$f(y_k) - f(x^*) \leq \frac{c_p L_p \|x^*\|^{p+1}}{k^{\frac{3p+1}{2}}}. \tag{18}$$

Furthermore, each iteration of ATD can be implemented in $\tilde{O}(1)$ calls to a p^{th}-order Taylor expansion oracle.

We first apply this general theorem to our smooth approximations, before showing overall higher-order smoothed oracle complexity for our non-smooth problems of interest. Here it will be useful to define $\hat{c}_p \stackrel{\text{def}}{=} 4^{2p-1} p (p+1)^{\frac{3p+1}{2}} \left(\frac{p+1}{\ln(p+2)}\right)^{p+1}$.

Corollaries 29 and 30, found in Appendix B, follow by combining Theorem 9 with Theorems 7 and 8, respectively. Thus, we may arrive at the following key theorems of this section (found in full in Appendix B), the proofs of which are immediate from the previous corollaries by using Fact 1 and Lemma 4.

Theorem 10 (Sketch) *Let $f(x) = \|\mathbf{A}x - b\|_\infty$ for $b \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times d}$ s.t. $\mathbf{A}^\top\mathbf{A} \succ 0$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies, for $N = \tilde{O}(1/\varepsilon^{(2p+2)/(3p+1)})$*

$$f(y_N) - f(x^*) \leq \varepsilon. \tag{19}$$

Theorem 11 (Sketch) *Let $f(x) = \lambda\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \max\{0, 1 - b_i\langle a_i, x\rangle\}$ where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ for $i \in [m]$, let $\tilde{\mathbf{Q}} \stackrel{def}{=} [b_1a_1\ b_2a_2\ \ldots\ b_ma_m]^\top$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies, for $N = \tilde{O}(1/\varepsilon^{(2p+2)/(3p+1)})$,*

$$f(y_N) - f(x^*) \leq \varepsilon.$$

4. Efficient implementation for $p = 3$

In this section, we go beyond the oracle model in the case of third-order smoothness (i.e., $p = 3$) in order to establish overall computational guarantees, beginning with ℓ_∞ regression:

Theorem 12 *Let $f(x) = \|\mathbf{A}x - b\|_\infty$ for $b \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times d}$ s.t. $\mathbf{A}^\top\mathbf{A} \succ 0$, and let $x^* \stackrel{def}{=} \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. There is a method, initialized with x_0, that outputs x_N such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

in $O\left(\frac{\log^{3/5}(m)\|x_0 - x^*\|^{4/5}_{\mathbf{A}^\top\mathbf{A}}}{\varepsilon^{4/5}}\right)$ iterations, where each iteration requires $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver, for some problem-dependent parameter $\mathcal{Z}$. [1]

Our results are also applicable to soft-margin SVMs, and so in particular, we get the following for ℓ_1-SVM (Bradley and Mangasarian, 1998; Zhu et al., 2004; Mangasarian, 2006).

Theorem 13 *Let $f(x) = \lambda\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \max\{0, 1 - b_i\langle a_i, x\rangle\}$ where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ for $i \in [m]$, let $\tilde{\mathbf{Q}} \stackrel{def}{=} [b_1a_1\ b_2a_2\ \ldots\ b_ma_m]^\top$, and let $x^* \stackrel{def}{=} \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. There is a method, initialized with x_0, that outputs x_N such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

in $O\left(\frac{(\lambda d)^{3/5}(\lambda d + \|\tilde{\mathbf{Q}}^\top\tilde{\mathbf{Q}}\|^2)^{1/5}\|x_0 - x^*\|^{4/5}}{\varepsilon^{4/5}}\right)$ iterations, where each iteration requires $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver, for some problem-dependent parameter $\mathcal{Z}$.

We begin by developing the necessary higher-order optimization guarantees, before later proving Theorems 12 and 13 in Appendix F.11. In order to handle the points that might be reached by our method, starting from an initial point x_0, we consider the following standard objects, beginning with the set

$$\mathscr{K} \stackrel{\text{def}}{=} \left\{x : \|x - x_0\|^2_{\mathbf{B}} \leq 4\|x_0 - x^*\|^2_{\mathbf{B}}\right\}. \tag{20}$$

1. $\mathcal{Z}$ depends polynomially upon, among other things, the diameter term $\mathcal{P}$ and the gradient norm bound $\mathcal{G}$—the full dependence may be found in the proof of Theorem 26. Note that $\mathcal{Z}$ only appears as part of polylogarithmic factors.

Given this set, we now consider the maximum function value attained over $\mathscr{K}$, i.e., $\mathcal{F} \stackrel{\text{def}}{=} \max_{x \in \mathscr{K}} f(x)$. Finally, we let

$$\mathcal{P} \stackrel{\text{def}}{=} \max_{x,y \in \mathscr{L}} \|x - y\|_{\mathbf{B}}^2, \tag{21}$$

where $\mathscr{L} \stackrel{\text{def}}{=} \{x : f(x) \leq \mathcal{F}\}$, and we let $\mathcal{G} \stackrel{\text{def}}{=} \max_{x \in \mathscr{L}} \|\nabla f(x)\|_{\mathbf{B}^{-1}}^2$.

We recall that

$$\Phi_{x,p}(y) \stackrel{\text{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} \nabla^i f(x)[y - x]^i, \; p \geq 1 \tag{22}$$

denotes the p^{th}-order Taylor approximation of $f(\cdot)$, centered at x. Furthermore, for $f(\cdot)$ that is (order p) L_p-smooth, we define a model function

$$\Omega_{x,p,\mathbf{B}}(y) \stackrel{\text{def}}{=} \Phi_{x,p}(y) + \frac{2pL_p}{(p+1)!} \|y - x\|_{\mathbf{B}}^{p+1}. \tag{23}$$

As we are only concerned in this section with functions that are third-order L_3-smooth, we will drop the p subscript to define $\Phi_x(y) \stackrel{\text{def}}{=} \Phi_{x,3}(y)$ and

$$\Omega_{x,\mathbf{B}}(y) \stackrel{\text{def}}{=} \Omega_{x,3,\mathbf{B}}(y) = \Phi_x(y) + \frac{L_3}{4} \|y - x\|_{\mathbf{B}}^4. \tag{24}$$

Note that $\Omega_{x,\mathbf{B}}(y)$ is third-order $6L_3$-smooth w.r.t $\|\cdot\|_{\mathbf{B}}$. The following theorem illustrates some useful properties of the model $\Omega_{x,\mathbf{B}}(\cdot)$.

Theorem 14 (Nesterov (2018a), Theorem 1, for $M = 2L_3$) *Suppose $f(\cdot)$ is convex, 3-times differentiable, and third-order L_3-smooth. Then, for any $x, y \in \mathbb{R}^d$, we have*

$$0 \preceq \nabla^2 f(y) \preceq \nabla^2 \Phi_x(y) + \frac{L_3}{2} \|y - x\|_{\mathbf{B}}^2 \mathbf{B}.$$

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

$$f(y) \leq \Omega_{x,\mathbf{B}}(y). \tag{25}$$

For functions $f(\cdot)$ that are third-order L_3-smooth w.r.t. $\|\cdot\|_{\mathbf{B}}$, we also have that, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(y) - \nabla \Phi_x(y)\|_{\mathbf{B}^{-1}} \leq \frac{L_3}{6} \|y - x\|_{\mathbf{B}}^3. \tag{26}$$

With this representation of the model function $\Omega_{x,\mathbf{B}}(\cdot)$ in hand, we let

$$T_{\mathbf{B}}(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{y \in \mathbb{R}^d} \Omega_{x,\mathbf{B}}(y) \tag{27}$$

denote a minimizer of the fourth-order model, centered at x. The following lemma concerning $\Omega_{x,\mathbf{B}}(\cdot)$ establishes a relaxed version of eq. (2.13) from Nesterov (2018a).

Lemma 15 *Let $\varepsilon > 0$, and let $T_{\mathbf{B}}(\cdot)$ be as in (27). Then, for all $x, y \in \mathbb{R}^d$,*

$$\langle \nabla f(x), y - x \rangle \geq \frac{1}{2L_3 \hat{r}_{\mathbf{B}}^2(x,y)} \|\nabla f(x)\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8} \hat{r}_{\mathbf{B}}^4(x,y) - \frac{2Z(x,y)W(x,y)\|x - T_{\mathbf{B}}(y)\|_{\mathbf{B}}}{2L_3 \hat{r}_{\mathbf{B}}^2(x,y)}, \tag{28}$$

for appropriately defined $Z(x,y)$, $W(x,y)$.

We may also observe that $\Omega_{x,\mathbf{B}}(\cdot)$ is (order 4) uniformly convex w.r.t. $\|\cdot\|_{\mathbf{B}}$.

Lemma 16 *For all $y, z \in \mathbb{R}^d$,*

$$\Omega_{x,\mathbf{B}}(z) \geq \Omega_{x,\mathbf{B}}(y) + \langle \nabla \Omega_{x,\mathbf{B}}(y), z - y \rangle + \frac{L_3}{12}\|z - y\|_{\mathbf{B}}^4. \tag{29}$$

4.1. Approximate auxiliary minimization

To begin, we consider the auxiliary minimization problem $\min_{h \in \mathbb{R}^d} \Gamma_{x,\mathbf{B}}(h)$, where

$$\Gamma_{x,\mathbf{B}}(h) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x) h + \frac{1}{6} \nabla^3 f(x)[h]^3 + \frac{L_3}{4}\|h\|_{\mathbf{B}}^4.$$

Note that $\Gamma_{x,\mathbf{B}}(h)$ is equivalent to $\Omega_{x,\mathbf{B}}(y)$, up to a change of variables. Our aim is to establish a minimization procedure which returns an $\tilde{\varepsilon}_{aam}$-optimal solution in $O(\log(\mathcal{A}/\tilde{\varepsilon}_{aam}))$ iterations, where $\mathcal{A}$ is defined in Corollary 18. Furthermore, each iteration is dominated by $O(\log^{O(1)}(1/\tilde{\varepsilon}_{aam}))$ calls to a linear system solver. This subroutine, which we call ApproxAuxMin (Algorithm 3), is described in Section 5 of Nesterov (2018a) and is able to return an approximate minimizer of $\Omega_{x,\mathbf{B}}(\cdot)$. The approach involves showing that the auxiliary function is relatively smooth and strongly convex (Bauschke et al., 2016; Lu et al., 2018), and further that each iteration of the method for minimizing such a function reduces to a minimization problem of the form

$$-\min_{\lambda > 0} w(\lambda), \tag{30}$$

where $w(\lambda) \stackrel{\text{def}}{=} \frac{\lambda^2}{2} + \frac{1}{2}\langle (\sqrt{2}\lambda \mathbf{B} + \nabla^2 f(x))^{-1} c_t, c_t \rangle$ and

$$c_t \stackrel{\text{def}}{=} \nabla \Gamma_{x,\mathbf{B}}(h_t) = \nabla f(x) + \nabla^2 f(x) h_t + \frac{1}{2}\nabla^3 f(x)[h_t]^2 + L_3 \|h_t\|_{\mathbf{B}}^2 \mathbf{B} h_t.$$

As noted by Nesterov (2018a), this minimization problem is both one-dimensional and strongly convex, and so we may achieve global linear convergence. Taken together with the relative smoothness and strong convexity of $\Gamma_{x,\mathbf{B}}(\cdot)$, we have the following theorem.

Theorem 17 (Nesterov (2018a), eq.(5.9) ($\tau = \sqrt{2}$). See also: Lu et al. (2018), Theorem 3.1) *For all h_t, $K \geq t \geq 0$, generated by* ApproxAuxMin$(y_k, \tilde{\varepsilon}_{aam})$ *(Algorithm 3), we have that*

$$\Gamma_{y_k,\mathbf{B}}(h_t) - \Gamma_{y_k,\mathbf{B}}(h^*) \leq \frac{\alpha}{\left(\frac{\sqrt{2}+1}{2}\right)^t - 1},$$

where $h^* \stackrel{def}{=} \operatorname{argmin}_{h \in \mathbb{R}^d} \Gamma_{y_k,\mathbf{B}}(h)$ and $\alpha \stackrel{def}{=} \frac{1}{\sqrt{2}}(h_0 - h^*)^\top \nabla^2 f(y_k)(h_0 - h^*) + \frac{\sqrt{2}L_3}{4}\|h_0 - h^*\|_{\mathbf{B}}^4$.

Corollary 18 *Let $x_{k+1} = y_k + h_K$ be the output from* ApproxAuxMin$(y_k, \tilde{\varepsilon}_{aam})$*, for $y_k \in \mathscr{L}$ and $K = O(\log(\mathcal{A}/\tilde{\varepsilon}_{aam}))$, where $\mathcal{A} \stackrel{def}{=} 1 + \max_{z \in \mathscr{L}} \frac{1}{\sqrt{2}}(T_{\mathbf{B}}(z) - z)^\top \nabla^2 f(z)(T_{\mathbf{B}}(z) - z) + \frac{\sqrt{2}L_3}{4}\|T_{\mathbf{B}}(z) - z\|_{\mathbf{B}}^4$. Then*

$$\Omega_{y_k,\mathbf{B}}(x_{k+1}) - \Omega_{y_k,\mathbf{B}}(T_{\mathbf{B}}(y_k)) \leq \tilde{\varepsilon}_{aam},$$

where each iteration requires time proportional to evaluating $f(\cdot)$ in order to compute c_t, as well as $O(\log^{O(1)}(1/\tilde{\varepsilon}_{aam}))$ calls to a linear system solver.

As we shall see, it will become necessary to handle the approximation error from ApproxAuxMin, and so we provide the following several lemmas to that end.

Lemma 19 *Let $\varepsilon > 0$, let x_{k+1} be as output by* $\mathsf{ApproxAuxMin}(y_k, \tilde{\varepsilon}_{aam})$*, and let $T_{\mathbf{B}}(y_k)$ be as in (27). Then,* $\|x_{k+1} - T_{\mathbf{B}}(y_k)\|_{\mathbf{B}} \leq \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}$.

Lemma 20 *Let* $x_{k+1} = \mathsf{ApproxAuxMin}(y_k, \tilde{\varepsilon}_{aam})$*. Then,*

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle \geq \frac{1}{2L_3 \hat{r}^2_{\mathbf{B}}(x_{k+1}, y_k)} \|\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} + \frac{3L_3}{8} \hat{r}^4_{\mathbf{B}}(x_{k+1}, y_k) - \frac{3Z(x_{k+1}, y_k) W(x_{k+1}, y_k) \tilde{\varepsilon}^{1/4}_{aam}}{L_3^{5/4} \hat{r}^2_{\mathbf{B}}(x_{k+1}, y_k)}.$$

Proof The result follows from Lemmas 15 and 19. ■

Lemma 21 *Let x_{k+1} be the output from* $\mathsf{ApproxAuxMin}(y_k, \tilde{\varepsilon}_{aam})$ *for $y_k \in \mathscr{L}$. In addition, let $r(y_k) \overset{def}{=} \|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}$. Then,*

$$\left|\hat{r}^2_{\mathbf{B}}(x_{k+1}, y_k) - r(y_k)^2\right| \leq 6 \left(\frac{\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4} \mathcal{P}^{1/2} + \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/2}.$$

4.2. Search procedure for finding ρ_k

In this section, we establish the correctness of RhoSearch (Algorithm 4), our subroutine for finding an appropriate choice of ρ_k, given x_k, v_k as inputs. One of the key algorithmic components for achieving fast higher-order acceleration, as observed by Monteiro and Svaiter (2013), is to determine ρ_k such that $\rho_k \approx \zeta_k(\rho_k)$, where we define

$$\zeta_k(\rho) \overset{\text{def}}{=} \|T_{\mathbf{B}}(y_k(\rho)) - y_k(\rho)\|^2_{\mathbf{B}}, \tag{31}$$

$$y_k(\rho) \overset{\text{def}}{=} (1 - \tau_k(\rho))x_k + \tau_k(\rho) v_k, \tag{32}$$

and

$$\tau_k(\rho) \overset{\text{def}}{=} \frac{2}{1 + \sqrt{1 + 4L_3 A_k \rho}}. \tag{33}$$

We will also need to define an approximate version

$$\hat{\zeta}_k(\rho) \overset{\text{def}}{=} \|x_{k+1}(\rho) - y_k(\rho)\|^2_{\mathbf{B}}, \tag{34}$$

where we let $x_{k+1}(\rho) \overset{\text{def}}{=} \mathsf{ApproxAuxMin}(y_k(\rho), \tilde{\varepsilon}_{aam})$. We may observe that $\zeta_k(\rho)$ is continuous in ρ, and furthermore that there exists some $0 \leq \rho^*_k \leq \infty$ such that $\zeta_k(\rho^*_k) = \rho^*_k$, since if $\rho = 0$, then $y_k = v_k$, and if $\rho_k \to \infty$, then $y_k = x_k$. Thus, we may reduce it to a binary search problem, under an appropriate initialization. For now, we assume that at each iteration $k \geq 0$, RhoSearch is given initial bounds ρ^-_{init} and ρ^+_{init} such that $\rho^-_{\text{init}} \leq \rho^*_k \leq \rho^+_{\text{init}}$, thus ensuring it is a valid binary search procedure. We will later show how FastQuartic can provide RhoSearch with such guarantees.

An important part of managing this process is to limit how quickly $\zeta_k(\rho)$ can grow, as we will need to ensure a closeness in function value once our candidate bounds ρ^- and ρ^+ are sufficiently close. Theorems 33 and 34, found in Appendix B, give us precisely what we need, namely a differential inequality w.r.t. $|\zeta'_k(\rho)|$. We note that the complicated description of $\zeta_k(\rho)$ as a function of ρ gives rise to several technical challenges.

4.3. Analyzing the convergence of FastQuartic

Algorithm 1 FastQuartic (Sketch)

Input: $x_0 = 0$, $A_0 = 0$, $\mathbf{B} \succ 0$, N.
Define $\psi_0(x) \overset{\text{def}}{=} \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2$.
for $k = 0$ **to** $N - 1$ **do**
  $v_k = \operatorname{argmin}_{x \in \mathbb{R}^d} \psi_k(x)$
  Find $\rho_k > 0$, $x_{k+1} \in \mathbb{R}^d$ such that $\rho_k \approx \|x_{k+1} - y_k\|_{\mathbf{B}}^2$, where:

$$a_{k+1} = \frac{1 + \sqrt{1 + 4L_3 A_k \rho_k}}{2L_3\rho_k} \qquad \left(\implies (a_{k+1})^2 = \frac{A_k + a_{k+1}}{L_3\rho_k} \right)$$
$$A_{k+1} = A_k + a_{k+1}, \qquad \tau_k = \frac{a_{k+1}}{A_{k+1}}, \qquad y_k = (1 - \tau_k)x_k + \tau_k v_k$$
$$x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^d} \Omega_{y_k, 3, \mathbf{B}}(x)$$

  $\psi_{k+1} = \psi_k + a_{k+1}\left[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle\right]$
end for
return x_N

Having shown the correctness of the binary search procedure in RhoSearch, we now describe our main algorithm, called FastQuartic (sketched in Algorithm 1), and prove its correctness. Due to space constraints, we include the full method (Algorithm 5), along with its subroutines, in Appendix E. Our analysis follows similarly to that of Chapter 4.3 in (Nesterov, 2018b), though we consider a higher-order model function for the case where $f(\cdot)$ is third-order L_3-smooth.

We begin by proving a useful inequality concerning the estimate sequence, which is a standard technique for analyzing accelerated methods (Nesterov, 2005b, 2018b). An important part of FastQuartic is to provide RhoSearch with appropriate ρ_{init}^+ and ρ_{init}^- that are valid upper and lower bounds, respectively, on ρ_k^*. As we will see, setting $\rho_{\text{init}}^+ = \mathcal{P}$ will provide a sufficiently large upper bound on ρ_k^*. For the lower bound, we will observe that, for a small enough choice of ρ_{init}^-, if it is still the case that $\rho_k^* < \rho_{\text{init}}^-$, then we can show that our current iterate achieves sufficiently small error, and so we are done. The following lemmas make these observations formal.

Lemma 22 *Let $c > 0$, $x_{k+1} = \text{ApproxAuxMin}(y_k, \tilde{\varepsilon}_{aam})$, where $y_k \in \mathscr{L}$, and suppose $c\rho_{init}^- \leq \hat{r}_{\mathbf{B}}^2(x_{k+1}, y_k)$. Then,*

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \geq \frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1}, y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1}, y_k) - \frac{\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{c\rho_{init}^-}.$$

where $\mathcal{W} > 0$ is some problem-dependent parameter.

Lemma 23 *For any $k \geq 0$, let A_k, x_k, v_k, $y_{i\{0 \leq i \leq k-1\}}$ be as generated by k iterations of* FastQuartic *with $\tilde{\varepsilon}_{aam} > 0$ chosen sufficiently small, and suppose that for all k iterations, $\rho_{init}^- \leq (1 + \tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ and $\rho_{init}^- \leq \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$ (for $\mathcal{Q}$ as in (45)). Then, we have that*

$$A_k f(x_k) + B_k \leq \psi_k^* \overset{def}{=} \min_{x \in \mathbb{R}^d} \psi_k(x), \tag{35}$$

where $B_k \overset{def}{=} \frac{3L_3}{16} \sum_{i=0}^{k-1} A_{i+1} \hat{r}_{\mathbf{B}}^4(x_{i+1}, y_i)$. *In addition,*

$$f(x_k) \leq \mathcal{F}, \quad \|v_k - x^*\|_{\mathbf{B}}^2 \leq \|x_0 - x^*\|_{\mathbf{B}}^2, \quad \text{and} \quad v_k, x_k \in \mathscr{L}. \tag{36}$$

Corollary 24 *For any* $k \geq 0$*, let* A_k*,* B_k*,* x_k *be as in the previous lemma statement. Then, we have*

$$f(x_k) - f(x^*) \leq \frac{1}{2A_k} \|x_0 - x^*\|_{\mathbf{B}}^2 \qquad \text{and} \qquad B_k \leq \frac{1}{2} \|x_0 - x^*\|_{\mathbf{B}}^2.$$

We now need to establish various bounds on the estimate sequence parameters A_k, namely Lemmas 35 and 36 which are found in Appendix B, again extending the analysis of Nesterov (2018b) to account for the higher-order smoothness. We thus arrive at the following key theorem.

Theorem 25 *Let* $k \geq 1$ *be such that the conditions in the statement of Lemma 23 hold. Then, we have*

$$f(x_k) - f(x^*) \leq \frac{128 L_3 \|x_0 - x^*\|_{\mathbf{B}}^4}{3} \left(\frac{2}{k+1} \right)^5. \tag{37}$$

So far, we have shown the correctness in the case where, for all $k \geq 0$, $\rho_{\text{init}}^- \leq (1 + \tilde{\varepsilon}_{fs}) \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ and $\rho_{\text{init}}^- \leq \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$. However, we need to ensure correctness of the case where, for some iteration of FastQuartic, it happens that $\rho_{\text{init}}^- > (1 + \tilde{\varepsilon}_{fs}) \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$, or $\rho_{\text{init}}^- \leq \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$. We handle these cases via Theorem 37 in Appendix B.

Having established the necessary results for proving the correctness of the output from ApproxAuxMin and RhoSearch, we may combine these observations with those of Section 4.3 to prove one of the key theorems of this work, which establishes the total cost of optimizing third-order smooth convex $f(\cdot)$. The proofs of Theorems 12 and 13 then follow, as discussed in Appendix F.11.

Theorem 26 *Suppose* $f(x)$ *is convex and third-order* L_3*-smooth. Then, under appropriate initialization,* FastQuartic *finds a point* x_N *such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

in $O\left(\left(\frac{L_3 \|x_0 - x^*\|_{\mathbf{B}}^4}{\varepsilon} \right)^{1/5} \right)$ *iterations, where each iteration requires* $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ *calls to a gradient oracle and linear system solver, and where* $\mathcal{Z}$ *is a polynomial in various problem-dependent parameters.*

Acknowledgments

We thank Deeksha Adil, Sébastien Bubeck, Yin Tat Lee, Sushant Sachdeva, Cyril Zhang, and Yi Zhang for numerous helpful conversations. We especially thank Richard Peng for his continued guidance and support in pursuit of this work. Parts of this work were done while BB was at Princeton University, where he was supported by Elad Hazan's NSF grant CCF-1704860.

References

Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for lp-norm regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1405–1424. SIAM, 2019a.

Deeksha Adil, Richard Peng, and Sushant Sachdeva. Fast, provably convergent irls algorithm for p-norm linear regression. In *Advances in Neural Information Processing Systems*, pages 14189–14200, 2019b.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

Amir Ali Ahmadi, Alex Olshevsky, Pablo A Parrilo, and John N Tsitsiklis. Np-hardness of deciding convexity of quartic polynomials and related problems. *Mathematical Programming*, 137(1-2): 453–476, 2013.

Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems*, pages 1614–1622, 2016.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, pages 1–34, 2018.

Michel Baes. Estimate sequence methods: extensions and approximations. 2009.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

Daniel Berend and Tamir Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Paul S Bradley and OL Mangasarian. Feature selection via concave minimization and support vector machines. In *International Conference on Machine Learning*, pages 82–90, 1998.

Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for l p regression provably beyond self-concordance and in input-sparsity time. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1130–1137. ACM, 2018a.

Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. *arXiv preprint arXiv:1812.08026*, 2018b.

Brian Bullins. Fast minimization of structured convex quartics. *arXiv preprint arXiv:1812.10349*, 2018.

Brian Bullins and Richard Peng. Higher-order accelerated methods for faster non-smooth optimization. *arXiv preprint arXiv:1906.01621*, 2019.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *arXiv preprint arXiv:2003.08078*, 2020.

Hui Han Chin, Aleksander Madry, Gary L Miller, and Richard Peng. Runtime guarantees for regression problems. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, pages 269–282. ACM, 2013.

Paul Christiano, Jonathan A Kelner, Aleksander Madry, Daniel A Spielman, and Shang-Hua Teng. Electrical flows, Laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, pages 273–282. ACM, 2011.

Michael B Cohen and Richard Peng. L p row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 183–192. ACM, 2015.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *arXiv preprint arXiv:1810.07896*, 2018.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

Alina Ene and Adrian Vladu. Improved convergence for ℓ_∞ and ℓ_1 regression via iteratively reweighted least squares. In *International Conference on Machine Learning*, 2019.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Dmitry Kovalev, Ahmed Mohhamed, Elena Chernousova, and César A. Uribe. The global rate of convergence for optimal tensor methods in smooth convex optimization. *arXiv preprint arXiv:1809.00382 (v10)*, 2018.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory*, pages 1392–1393, 2019.

Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. *arXiv preprint arXiv:1812.06557*, 2018.

Jonathan A Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 217–226. SIAM, 2014.

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 *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433, 2014. doi: 10.1109/FOCS.2014.52. URL https://doi.org/10.1109/FOCS.2014.52.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Olvi L Mangasarian. Exact 1-norm support vector machines via unconstrained convex differentiable minimization. *Journal of Machine Learning Research*, 7(Jul):1517–1530, 2006.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005a.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005b.

Yu Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2018a.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2018b.

Pablo A Parrilo and Bernd Sturmfels. Minimizing polynomial functions. In *Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 60, pages 83–99, 2003.

John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, April 1998.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.

Jonah Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 452–460. ACM, 2017.

Aaron Sidford and Kevin Tian. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science*, pages 922–933, 2018.

Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in Neural Information Processing Systems*, pages 49–56, 2004.

Appendix A. Related work

Smooth approximation techniques: It has been shown that one can go beyond the black-box convergence of $O(1/\varepsilon^2)$ to achieve an $O(1/\varepsilon)$ rate for certain classes of non-smooth functions (Nemirovski, 2004; Nesterov, 2005b,a, 2007). One such approach by Nesterov (2005b) was to carefully smooth the well-structured function, and the work goes on to present several applications of the method, including ℓ_∞ and ℓ_1 regression, in addition to saddle-point games. However, the methods for all of these examples incur an $O(1/\varepsilon)$ dependence which remains in several works that build upon these techniques (Sherman, 2017; Sidford and Tian, 2018). For a more comprehensive overview, we refer the reader to (Beck and Teboulle, 2012).

Higher-order accelerated methods: Several works have considered accelerated variants of optimization methods based on access to higher-order derivative information. Nesterov (2008) showed that one can accelerate cubic regularization, under a Lipschitz Hessian condition, to attain faster convergence, and these results were later generalized by Baes (2009) to arbitrary higher-order oracle access under the appropriate notions of higher-order smoothness. The rate attained in (Nesterov, 2008) was further improved upon by Monteiro and Svaiter (2013), and lower bounds have established that the oracle complexity of this result is nearly tight (up to logarithmic factors) when the Hessian is Lipschitz continuous (Arjevani et al., 2018). Until recently, however, it was an open question whether these lower bounds are tight for general higher-order oracle access (and smoothness), though this question has been mostly resolved as a result of several works developed over the past year (Gasnikov et al., 2018; Jiang et al., 2018; Bubeck et al., 2018b; Bullins, 2018; Gasnikov et al., 2019).

ℓ_∞ regression: Various regression problems play a central role in numerous computational and learning tasks. Designing better methods for ℓ_∞ regression in particular has led to faster approximate max flow algorithms (Christiano et al., 2011; Chin et al., 2013; Kelner et al., 2014; Sherman, 2017; Sidford and Tian, 2018). Recently, Ene and Vladu (2019) presented a method for ℓ_∞ regression, based on iteratively reweighted least squares, that achieves an iteration complexity of $O(m^{1/3}\log(1/\varepsilon)/\varepsilon^{2/3} + \log(m/\varepsilon)/\varepsilon^2)$. We note that their rate of convergence has an $O(m^{1/3})$ dependence, whereas our result (Theorem 12) only depends logarithmically in m, though with an additional diameter dependence, i.e., $\|x_0 - x^*\|^{4/5}$.

Soft-margin SVM: Support vector machines (SVMs) (Cortes and Vapnik, 1995) have enjoyed widespread adoption for classification tasks in machine learning (Cristianini et al., 2000). For the soft-margin version, several approaches have been proposed for dealing with the non-smooth nature

of the hinge loss. The standard approach is to cast the (ℓ_2-regularized) SVM problem as a quadratic programming problem (Platt, 1998; Boyd and Vandenberghe, 2004). Stochastic sub-gradient methods have also been successful due to their advantage in per-iteration cost (Shalev-Shwartz et al., 2011). While ℓ_2-SVM is arguably the most well-known variant, ℓ_p-SVMs, for general $p \geq 1$, have also been studied (Bradley and Mangasarian, 1998). ℓ_1-SVMs (Zhu et al., 2004; Mangasarian, 2006) are appealing, in particular, due to their sparcity-inducing tendencies, though they forfeit the strong convexity guarantees that come with ℓ_2 regularization (Allen-Zhu and Hazan, 2016).

Interior-point methods: It is well-known that both ℓ_∞ regression and ℓ_1-SVM can be expressed as linear programs (Boyd and Vandenberghe, 2004; Bradley and Mangasarian, 1998), and thus are amenable to fast LP solvers (Lee and Sidford, 2014; Cohen et al., 2018). In particular, this means that each can be solved in either $\tilde{O}(d^\omega)$ time (where $\omega \sim 2.373$ is the matrix multiplication constant) (Cohen et al., 2018), or in $\tilde{O}(\sqrt{d})$ linear system solves (Lee and Sidford, 2014). We note that, while these methods dominate in the high-accuracy regime, our method is competitive, under modest choices of ε and favorable linear system solves, when $\|x_0 - x^*\|^{4/5} \leq O(\sqrt{d})$ (up to logarithmic factors).

Appendix B. Additional theorems

Corollary 27 *Let $f_\mu(x) = \operatorname{smax}_\mu(\mathbf{A}x - b)$ be the softmax approximation to (6) for $\mu = \frac{\varepsilon}{2\log(m)}$, where $\mathbf{A}$ is such that $\mathbf{A}^\top\mathbf{A} \succ 0$. Then, letting $x_\mu^* \overset{def}{=} \operatorname{argmin}_{x\in\mathbb{R}^d} f_\mu(x)$, FastQuartic finds a point x_N such that*

$$f_\mu(x_N) - f_\mu(x_\mu^*) \leq \frac{\varepsilon}{2}$$

in $O\left(\frac{\log^{3/5}(m)\|x_0 - x^*\|^{4/5}_{\mathbf{A}^\top\mathbf{A}}}{\varepsilon^{4/5}}\right)$ iterations, where each iteration requires $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver, and where $\mathcal{Z}$ is a polynomial in various problem-dependent parameters.

Corollary 28 *Let $f_\mu(x) = \lambda\,\text{soft-}\ell1_\mu(x) + \text{softSVM}_\mu(\tilde{\mathbf{Q}}x)$ be the smooth approximation to $f(x)$ (as in (7)) with $\mu = \frac{\varepsilon}{4\lambda d}$ for $\varepsilon > 0$. Then, letting $x_\mu^* \overset{def}{=} \operatorname{argmin}_{x\in\mathbb{R}^d} f_\mu(x)$, FastQuartic finds a point x_N such that*

$$f_\mu(x_N) - f_\mu(x_\mu^*) \leq \frac{\varepsilon}{2}$$

in $O\left(\frac{(\lambda d)^{3/5}(\lambda d + \|\tilde{\mathbf{Q}}^\top\tilde{\mathbf{Q}}\|^2)^{1/5}\|x_0 - x^*\|^{4/5}}{\varepsilon^{4/5}}\right)$ iterations, where each iteration requires $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver, and where $\mathcal{Z}$ is a polynomial in various problem-dependent parameters.

Corollary 29 *Let $f_\mu(x) = \operatorname{smax}_\mu(\mathbf{A}x - b)$ be the softmax approximation to (6) for $\mu = \frac{\varepsilon}{2\log(m)}$, where $\mathbf{A}$ is such that $\mathbf{A}^\top\mathbf{A} \succ 0$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies*

$$f_\mu(y_N) - f(x^*) \leq \frac{\varepsilon}{2} \tag{38}$$

for $N = \left\lceil \left(\frac{\hat{c}_p\|x^*\|^{p+1}_{\mathbf{A}^\top\mathbf{A}}\log^p(m)}{\varepsilon^{p+1}}\right)^{\frac{2}{3p+1}} \right\rceil$.

Corollary 30 *Let $\varepsilon > 0$, let $f_\mu(x) = \lambda\,\text{soft-}\ell 1_\mu(x) + \text{softSVM}(\tilde{\mathbf{Q}}x)$ be the softmax approximation to (7) for $\mu = \frac{\varepsilon}{4\lambda d}$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies*

$$f_\mu(y_N) - f(x^*) \leq \frac{\varepsilon}{2} \tag{39}$$

for $N = \left\lceil \left(\frac{\hat{c}_p \|x^*\|^{p+1} (\lambda d)^p (\lambda d + \|\tilde{\mathbf{Q}}^\top \tilde{\mathbf{Q}}\|^{\frac{p+1}{2}})}{\varepsilon^{p+1}} \right)^{\frac{2}{3p+1}} \right\rceil$.

Theorem 31 *Let $f(x) = \|\mathbf{A}x - b\|_\infty$ for $b \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times d}$ s.t. $\mathbf{A}^\top \mathbf{A} \succ 0$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies*

$$f(y_N) - f(x^*) \leq \varepsilon \tag{40}$$

for $N = \left\lceil \left(\frac{\hat{c}_p \|x^*\|^{p+1}_{\mathbf{A}^\top \mathbf{A}} \log^p(m)}{\varepsilon^{p+1}} \right)^{\frac{2}{3p+1}} \right\rceil$.

Theorem 32 *Let $f(x) = \lambda\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \max\{0, 1 - b_i\langle a_i, x\rangle\}$ where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ for $i \in [m]$, let $\tilde{\mathbf{Q}} \overset{def}{=} [b_1 a_1\ b_2 a_2\ \ldots\ b_m a_m]^\top$, and let x^* denote a minimizer of $f(\cdot)$. Then, ATD satisifies*

$$f(y_N) - f(x^*) \leq \varepsilon$$

for $N = \left\lceil \left(\frac{\hat{c}_p \|x^*\|^{p+1} (\lambda d)^p (\lambda d + \|\tilde{\mathbf{Q}}^\top \tilde{\mathbf{Q}}\|^{\frac{p+1}{2}})}{\varepsilon^{p+1}} \right)^{\frac{2}{3p+1}} \right\rceil$.

Theorem 33 *Let $\zeta_k(\rho) > 0$ be as defined in (31), for some $y_k(\rho) \in \mathscr{L}$. Then we have that, for all $\rho \geq \rho_{init}^-$,*

$$\left|\zeta_k'(\rho)\right| \leq \frac{\mathcal{R}}{\zeta_k(\rho)^{1/2}},$$

where $\mathcal{R}$ is as defined in (43).

Proof Note that $\zeta_k(\rho) = (m \circ y_k)(\rho)$, where $m(y_k) = \|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2$ and $y_k(\rho)$ is as defined in (32). Therefore, by the chain rule, we have

$$\begin{aligned}
\left|\zeta_k'(\rho)\right| &= |\mathbf{J}_\rho y_k(\rho) \nabla_{y_k} m(y_k(\rho))| \\
&\leq \|\mathbf{J}_\rho y_k(\rho)\|_{\mathbf{B}} \|\nabla_{y_k} m(y_k(\rho))\|_{\mathbf{B}^{-1}} \\
&\leq \lambda_{\max}(\mathbf{B}^{-1})^{1/2} \|\mathbf{J}_\rho y_k(\rho)\|_{\mathbf{B}} \|\nabla_{y_k} m(y_k(\rho))\|,
\end{aligned}$$

where we let $\mathbf{J}$ denote the Jacobian. For $\|\mathbf{J}_\rho y_k(\rho)\|_{\mathbf{B}}$, we know by (32) and (33) that

$$y_k(\rho) = (1 - \tau_k(\rho))x_k + \tau_k(\rho) v_k$$

and

$$\tau_k(\rho) = \frac{2}{1 + \sqrt{1 + 4L_3 A_k \rho}}.$$

Thus, it follows that

$$\mathbf{J}_\rho y_k(\rho) = -\frac{d}{d\rho}\tau_k(\rho)\cdot x_k + \frac{d}{d\rho}\tau_k(\rho)\cdot v_k.$$

Note that

$$\left|\frac{d}{d\rho}\tau_k(\rho)\right| = \frac{4L_3A_k}{(1+\sqrt{1+4L_3A_k\rho})^2\sqrt{1+4L_3A_k\rho}} \leq \frac{4L_3A_k}{(1+4L_3A_k\rho)^{3/2}} \leq \frac{1}{\rho} \quad .$$

Taken together, this gives us that

$$\|\mathbf{J}_\rho y_k(\rho)\|_{\mathbf{B}} \leq \left|\frac{d}{d\rho}\tau_k(\rho)\right| (\|x_k\|_{\mathbf{B}} + \|v_k\|_{\mathbf{B}}) \leq \frac{\|x_k\|_{\mathbf{B}} + \|v_k\|_{\mathbf{B}}}{\rho}.$$

To provide a bound for $\|\nabla_{y_k} m(y_k(\rho))\|$, we begin by letting $g(x,z) \overset{\text{def}}{=} \Omega_{x,\mathbf{B}}(z)$. We may see that $T_{\mathbf{B}}(y_k) = \operatorname{argmin}_{z\in\mathbb{R}^d} g(y_k,z)$. As long as $\left[\partial_z^2 g(y_k, T_{\mathbf{B}}(y_k))\right]^{-1} \succ 0$, which we will see holds when $\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}} > 0$, we have that, by the implicit function theorem,

$$\mathbf{J}_x T_{\mathbf{B}}(x) = -\left[\partial_z^2 g(x, T_{\mathbf{B}}(x))\right]^{-1} \partial_x\partial_z g(x, T_{\mathbf{B}}(x)).$$

Note that, since $g(x,z) = \Phi_x(z) + \frac{L_3}{4}\|z-x\|_{\mathbf{B}}^4$, we have

$$\partial_z g(x,z) = \nabla f(x) + \nabla^2 f(x)[z-x] + \frac{1}{2}\nabla^3 f(x)[z-x]^2 + L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B}(z-x),$$

and so it follows that

$$\begin{aligned}\partial_z^2 g(x,z) &= \nabla^2 f(x) + \nabla^3 f(x)[z-x] + 2L_3\mathbf{B}(z-x)(z-x)^\top\mathbf{B} + L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B} \\ &\succeq \nabla^2 f(x) + \nabla^3 f(x)[z-x] + L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B},\end{aligned}$$

and

$$\begin{aligned}\partial_x\partial_z g(x,z) &= \nabla^2 f(x) + \nabla^3 f(x)[z-x] - \nabla^2 f(x) + \frac{1}{2}\nabla^4 f(x)[z-x]^2 \\ &\quad - \nabla^3 f(x)[z-x] + 2L_3\mathbf{B}(z-x)(z-x)^\top\mathbf{B} - L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B} \\ &= \nabla^4 f(x)[z-x]^2 + 2L_3\mathbf{B}(z-x)(z-x)^\top\mathbf{B} - L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B}.\end{aligned}$$

Thus,

$$\|\partial_x\partial_z g(x,z)\| \leq H(x,z), \tag{41}$$

where

$$H(x,z) \overset{\text{def}}{=} \|\nabla^4 f(x)[z-x]^2\| + 2L_3\|\mathbf{B}(z-x)(z-x)^\top\mathbf{B}\| + L_3\|z-x\|_{\mathbf{B}}^2\|\mathbf{B}\|.$$

By Theorem 14 we have that $\nabla^2 f(x) + \nabla^3 f(x)[z-x] + \frac{L_3}{2}\|z-x\|_{\mathbf{B}}^2\mathbf{B} \succeq 0$, and so

$$\begin{aligned}\partial_z^2 g(x,z) &\succeq \nabla^2 f(x) + \nabla^3 f(x)[z-x] + L_3\|z-x\|_{\mathbf{B}}^2\mathbf{B} \\ &\succeq \frac{L_3\|z-x\|_{\mathbf{B}}^2}{2}\mathbf{B}.\end{aligned}$$

Thus,

$$\|\left[\partial_z^2 g(x,z)\right]^{-1}\| \leq \frac{1}{\lambda_{\min}\left([\partial_z^2 g(x,z))]\right)} \leq \frac{2}{L_3 \lambda_{\min}(\mathbf{B})\|z-x\|_{\mathbf{B}}^2}. \tag{42}$$

We may now observe that, for $m(y)$,

$$\nabla_{y_k} m(y_k) = 2(\mathbf{J}_{y_k} T(y_k) - \mathbf{I})\mathbf{B}(T(y_k) - y_k),$$

and so, by standard matrix norm inequalities,

$$\begin{aligned}
\|\nabla_{y_k} m(y_k)\| &= 2\|(\mathbf{J}_{y_k} T_{\mathbf{B}}(y_k) - \mathbf{I})\mathbf{B}^{1/2}\mathbf{B}^{1/2}(T(y_k) - y_k)\| \\
&\leq 2\|\mathbf{J}_{y_k} T_{\mathbf{B}}(y_k)\| \cdot \|\mathbf{B}^{1/2}\| \cdot \|T(y_k) - y_k\|_{\mathbf{B}} + \|\mathbf{B}^{1/2}\| \cdot \|T(y_k) - y_k\|_{\mathbf{B}} \\
&\leq 2\lambda_{\max}(\mathbf{B}^{1/2}) \\
&\quad \cdot \left(\|\left[\partial_z^2 g(y_k, T_{\mathbf{B}}(y_k))\right]^{-1} \partial_x \partial_z g(y_k, T_{\mathbf{B}}(y_k))\| \cdot \|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}} + \|T(y_k) - y_k\|_{\mathbf{B}}\right) \\
&\leq 2\lambda_{\max}(\mathbf{B}^{1/2}) \\
&\quad \cdot \left(\|\left[\partial_z^2 g(y_k, T_{\mathbf{B}}(y_k))\right]^{-1}\| \cdot \|\partial_x \partial_z g(y_k, T_{\mathbf{B}}(y_k))\| \cdot \|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}} + \|T(y_k) - y_k\|_{\mathbf{B}}\right) \\
&\leq 2\lambda_{\max}(\mathbf{B}^{1/2}) \left(\frac{2H(y_k, T_{\mathbf{B}}(y_k)) + L_3 \lambda_{\min}(\mathbf{B})\|T(y_k) - y_k\|_{\mathbf{B}}^2}{L_3 \lambda_{\min}(\mathbf{B})\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}}\right)
\end{aligned}$$

where the last inequality follows from (41) and (42), and since $\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}} > 0$ (as if $T_{\mathbf{B}}(y_k) = y_k$, then y_k is a minimizer of $f(\cdot)$).

All together, this gives us that

$$\begin{aligned}
\left|\zeta'(\rho)\right| &\leq \lambda_{\max}(\mathbf{B}^{-1})^{1/2}\|\mathbf{J}_\rho y_k(\rho)\|\|\nabla_{y_k} m(y_k(\rho))\| \\
&\leq \lambda_{\max}(\mathbf{B}^{-1})^{1/2} \left(\frac{\|x_k\|_{\mathbf{B}} + \|v_k\|_{\mathbf{B}}}{\rho}\right) \\
&\quad \cdot \left(2\lambda_{\max}(\mathbf{B}^{1/2}) \left(\frac{2H(y_k(\rho), T_{\mathbf{B}}(y_k(\rho))) + L_3 \lambda_{\min}(\mathbf{B})\|T(y_k(\rho)) - y_k(\rho)\|_{\mathbf{B}}^2}{L_3 \lambda_{\min}(\mathbf{B})\|T_{\mathbf{B}}(y_k(\rho)) - y_k(\rho)\|_{\mathbf{B}}}\right)\right).
\end{aligned}$$

Let $\mathcal{H} \overset{\text{def}}{=} \max_{x,z \in \mathscr{L}} H(x,z)$, ρ_{init}^- be our initial lower bound on ρ_k^*, and $\mathcal{P}$ be as in (21). Since $y_k(\rho) \in \mathscr{L}$ and $\zeta(\rho) = \|T_{\mathbf{B}}(y_k(\rho)) - y_k(\rho)\|_{\mathbf{B}}^2$ by definition, it follows that

$$\left|\zeta'(\rho)\right| \leq \frac{\mathcal{R}}{\zeta(\rho)^{1/2}},$$

where

$$\mathcal{R} \overset{\text{def}}{=} \frac{4\mathcal{P}^{1/2}\lambda_{\max}(\mathbf{B}^{1/2})\left(2\mathcal{H} + L_3 \lambda_{\min}(\mathbf{B})\mathcal{P}\right)}{L_3 \lambda_{\min}(\mathbf{B})\rho_{\text{init}}^-}. \tag{43}$$

■

With this differential inequality in hand, we may now provide an important approximation guarantee for ρ_k.

Theorem 34 *Given $x_k, v_k \in \mathscr{L}$, $0 < \tilde{\varepsilon}_{rs} < 1$ as inputs, and $\tilde{\varepsilon}_{aam} > 0$ chosen sufficiently small, the* RhoSearch *algorithm outputs ρ_k and x_{k+1} such that*

$$(1 - \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho_k) \leq \rho_k \leq (1 + \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho_k) \tag{44}$$

where $\hat{\zeta}_k(\cdot)$ is as defined in (34).

Proof By sufficiently small, we mean that $\tilde{\varepsilon}_{aam}$ is chosen such that

$$\tilde{\varepsilon}_{aam} \leq \min\left\{\left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{Q}}\right)^4, \left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{W}}\right)^4, \frac{1}{2}\right\},$$

for $\mathcal{W}$ as defined in (65), and for

$$\mathcal{Q} \stackrel{\text{def}}{=} \left(\frac{6\mathcal{P}^{1/2}}{L_3^{1/4}} + \frac{5}{L_3^{1/2}}\right). \tag{45}$$

We proceed by proving the correctness of the binary search procedure. Consider $\hat{\rho}$ from the algorithm, and let $\hat{x}_{k+1}$ be the output from the call to $\mathsf{ApproxAuxMin}(\hat{y}_k, \tilde{\varepsilon}_{aam})$ in the RhoSearch algorithm. Then, at each iteration, one of the following three conditions must hold:

$$\text{(a) } \hat{\rho} > \hat{\zeta}_k(\hat{\rho}) + \tilde{\delta}; \text{ or}$$

$$\text{(b) } \hat{\rho} < \hat{\zeta}_k(\hat{\rho}) - \tilde{\delta}; \text{ or}$$

$$\text{(c) } \hat{\zeta}_k(\hat{\rho}) - \tilde{\delta} \leq \hat{\rho} \leq \hat{\zeta}_k(\hat{\rho}) + \tilde{\delta},$$

where

$$\tilde{\delta} \stackrel{\text{def}}{=} 6\left(\frac{\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}\mathcal{P}^{1/2} + \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/2}.$$

Note that, based on our choice of $\tilde{\varepsilon}_{aam}$, we ensure that $\tilde{\delta} \leq \frac{\tilde{\varepsilon}_{rs}^2}{4}$. Suppose condition (a) holds. Then, by Lemma 21 (with $y_k = y_k(\hat{\rho})$), we have that $\zeta_k(\hat{\rho}) - \tilde{\delta} \leq \hat{\zeta}(\hat{\rho})$, and so it follows that $\hat{\rho} > \zeta_k(\hat{\rho})$. Thus, $\hat{\rho}$ is an upper bound on ρ_k^*, and so this proves the correctness ρ^+ remaining an upper bound on ρ_k^* after updating $\rho^+ \leftarrow \hat{\rho}$. By similar reasoning, we may conclude that if condition (b) holds, $\hat{\rho}$ is a lower bound on ρ_k^*, and so ρ^- remains a lower bound on ρ_k^* after updating $\rho^- \leftarrow \hat{\rho}$.

If condition (c) holds, then it must be the case that $\hat{\zeta}_k(\hat{\rho}) \geq \frac{\tilde{\varepsilon}_{rs}}{2}$, since if we suppose that $\hat{\zeta}_k(\hat{\rho}) < \frac{\tilde{\varepsilon}_{rs}}{2}$, this implies that $\hat{\rho} \leq \hat{\zeta}_k(\hat{\rho}) + \tilde{\delta} \leq \frac{3\tilde{\varepsilon}_{rs}}{4}$. However, this is a contradicition since we ensure that $\hat{\rho} \geq \rho_{\text{init}}^- \geq \tilde{\varepsilon}_{rs}$. Therefore, since $\tilde{\delta} \leq \frac{\tilde{\varepsilon}_{rs}^2}{4} \leq \tilde{\varepsilon}_{rs}\hat{\zeta}_k(\hat{\rho})$, it follows that

$$(1 - \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\hat{\rho}) \leq \hat{\rho} \leq (1 + \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\hat{\rho}),$$

which means that condition (44) is met.

Based on our choice of update, anytime condition (a) or (b) holds and the update takes place, we guarantee a decrease in $|\rho^+ - \rho^-|$, and so after $O(\log(\mathcal{R}/\tilde{\varepsilon}_{rs}))$ iterations, we are assured that $|\rho^+ - \rho^-| \leq \frac{\tilde{\varepsilon}_{rs}^3}{100\mathcal{R}}$. At this point, we make use of Theorem 33 to argue that ρ^- must fall in the

desired range, i.e., $(1-\tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho^-) \leq \rho^- \leq (1+\tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho^-)$. To show this, we first note that $|\rho_k^* - \rho^-| \leq \frac{\tilde{\varepsilon}_{rs}^3}{100\mathcal{R}}$. Thus, using the fact that $\zeta_k(\rho) \geq 0$, Theorem 33 implies that

$$\left|\zeta_k'(\rho)(\zeta_k(\rho))^{1/2}\right| \leq \mathcal{R} \implies -\mathcal{R} \leq \zeta_k'(\rho)(\zeta_k(\rho))^{1/2} \leq \mathcal{R}.$$

Note that $\rho^- \leq \rho_k^*$. By integrating with respect to ρ, we have

$$\int_{\rho_k^*}^{\rho^-} \mathcal{R}d\rho \leq \int_{\rho_k^*}^{\rho^-} \zeta_k'(\rho)(\zeta_k(\rho))^{1/2}d\rho \leq \int_{\rho_k^*}^{\rho^-} -\mathcal{R}d\rho.$$

It follows that

$$\frac{2}{3}\zeta_k(\rho_k^*)^{3/2} + \mathcal{R}(\rho^- - \rho_k^*) \leq \frac{2}{3}\zeta_k(\rho^-)^{3/2} \leq \frac{2}{3}\zeta_k(\rho_k^*)^{3/2} - \mathcal{R}(\rho^- - \rho_k^*),$$

and so we have

$$\left(\zeta_k(\rho_k^*)^{3/2} + \frac{3\mathcal{R}}{2}(\rho^- - \rho_k^*)\right)^{2/3} \leq \zeta_k(\rho^-) \leq \left(\zeta_k(\rho_k^*)^{3/2} - \frac{3\mathcal{R}}{2}(\rho^- - \rho_k^*)\right)^{2/3}.$$

We may now observe that

$$\begin{aligned}
\zeta_k(\rho^-) &\leq \left(\zeta_k(\rho_k^*)^{3/2} - \frac{3\mathcal{R}}{2}(\rho^- - \rho_k^*)\right)^{2/3} \\
&= \left(\zeta_k(\rho_k^*)^{3/2} + \frac{3\mathcal{R}}{2}(\rho_k^* - \rho^-)\right)^{2/3} \\
&\leq \left(\zeta_k(\rho_k^*)^{3/2} + \frac{\tilde{\varepsilon}_{rs}^3}{50}\right)^{2/3} \\
&\leq \zeta_k(\rho_k^*) + \left(\frac{1}{50}\right)^{2/3}\tilde{\varepsilon}_{rs}^2,
\end{aligned}$$

and so

$$\zeta_k(\rho^-) - \zeta_k(\rho_k^*) \leq \frac{\tilde{\varepsilon}_{rs}^2}{10}. \tag{46}$$

We again use Lemma 21 to see that

$$\left|\zeta(\rho^-) - \hat{\zeta}_k(\rho^-)\right| \leq 6\left(\frac{\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}\mathcal{P}^{1/2} + \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/2} \leq \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}, \tag{47}$$

where $\mathcal{Q}$ is as defined in (45),

and the last inequality follows from the fact that $\tilde{\varepsilon}_{aam} \leq \frac{1}{2}$. Thus, since by our choice of $\tilde{\varepsilon}_{aam}$ we know that $\tilde{\varepsilon}_{aam} \leq \left(\frac{\tilde{\varepsilon}_{rs}^2}{100\mathcal{Q}}\right)^4$, it follows that

$$\left|\zeta_k(\rho^-) - \hat{\zeta}_k(\rho^-)\right| \leq \frac{\tilde{\varepsilon}_{rs}^2}{100}.$$

For the sake of clarity, we assume $\mathcal{R} \geq 1$ – otherwise, we can choose $M = O(\log(1/\tilde{\varepsilon}_{rs}))$, and a similar analysis holds. Taken together with (46) and the fact that $|\rho^- - \rho_k^*| \leq \frac{\tilde{\varepsilon}_{rs}^3}{100\mathcal{R}}$ and $\tilde{\varepsilon}_{rs} \leq 1$, we have that

$$\rho^- \geq \rho_k^* - \frac{\tilde{\varepsilon}_{rs}^3}{100\mathcal{R}} = \zeta_k(\rho_k^*) - \frac{\tilde{\varepsilon}_{rs}^3}{100\mathcal{R}} \geq \zeta_k(\rho_k^*) - \frac{\tilde{\varepsilon}_{rs}^2}{100} \geq \zeta_k(\rho^-) - \frac{11\tilde{\varepsilon}_{rs}^2}{100} \geq \hat{\zeta}_k(\rho_k^-) - \frac{12\tilde{\varepsilon}_{rs}^2}{100}.$$

Note that, by a similar reasoning as above, it must be the case that $\hat{\zeta}_k(\rho^-) \geq \frac{\tilde{\varepsilon}_{rs}}{2}$. Since we have ensured throughout the procedure that $\rho^- \leq \hat{\zeta}_k(\rho^-)$, it follows that

$$(1 - \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho^-) \leq \rho^- \leq (1 + \tilde{\varepsilon}_{rs})\hat{\zeta}_k(\rho^-),$$

as desired, and so we set $\rho_k = \rho^-$. ■

Lemma 35 *For any $k \geq 0$, we have that*

$$A_k \geq \frac{1}{4L_3}\left(\sum_{i=0}^{k-1} \frac{1}{\rho_i^{1/2}}\right)^2,$$

and thus $A_k \geq \frac{1}{4L_3\rho_i}$, for all $i \in \{0, \ldots, k-1\}$.

Proof Note that, by our choice of A_k and a_k,

$$A_{k+1}^{1/2} - A_k^{1/2} = \frac{a_{k+1}}{A_{k+1}^{1/2} + A_k^{1/2}} = \frac{1}{A_{k+1}^{1/2} + A_k^{1/2}}\sqrt{\frac{A_{k+1}}{L_3\rho_k}} \geq \sqrt{\frac{1}{4L_3\rho_k}}. \tag{48}$$

Again, we procede with a proof by induction. $A_0 = 0$, thus the case for $k = 0$ holds. Now, suppose for some $k \geq 0$,

$$A_k \geq \frac{1}{4L_3}\left(\sum_{i=0}^{k-1} \frac{1}{\rho_i^{1/2}}\right)^2.$$

By (48), we know that

$$A_{k+1}^{1/2} \geq A_k^{1/2} + \sqrt{\frac{1}{4L_3\rho_k}} \geq \sqrt{\frac{1}{4L_3}}\sum_{i=0}^{k-1}\frac{1}{\rho_i^{1/2}} + \sqrt{\frac{1}{4L_3\rho_k}} = \sqrt{\frac{1}{4L_3}}\sum_{i=0}^{k}\frac{1}{\rho_i^{1/2}}$$

which concludes the induction step. ■

Lemma 36 *For any $k \geq 1$, we have*

$$A_k \geq \frac{3}{256L_3\|x_0 - x^*\|_{\mathbf{B}}^2}\left(\frac{k+1}{2}\right)^5. \tag{49}$$

Proof Using Theorem 34 and the fact that $\tilde{\varepsilon}_{rs} < 1$, we have that $\rho_i \leq 2\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)$. By Lemma 35, it follows that, for all $k \geq 0$,

$$A_k \geq \frac{1}{4L_3}\left(\sum_{i=0}^{k-1}\frac{1}{\rho_i^{1/2}}\right)^2 \geq \frac{1}{8L_3}\left(\sum_{i=0}^{k-1}\frac{1}{\hat{r}_{\mathbf{B}}(x_{i+1}, y_i)}\right)^2. \tag{50}$$

Note that, for all $k \geq 0$, $x \in \mathbb{R}^d$,

$$\begin{aligned}\psi_k(x) &= \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \sum_{i=0}^{k} a_i\left[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle\right] \\ &\leq \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \sum_{i=0}^{k} a_i f(x) \\ &= A_k f(x) + \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}},\end{aligned}$$

and so it follows that

$$A_k f(x_k) + B_k \leq \min_{x\in\mathbb{R}^d}\psi_k(x) \leq \min_{x\in\mathbb{R}^d} A_k f(x) + \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} = A_k f(x^*) + \frac{1}{2}\|x^* - x_0\|^2_{\mathbf{B}}.$$

Rearranging, we have

$$\frac{3L_3}{16}\sum_{i=0}^{k-1} A_{i+1}\hat{r}^4_{\mathbf{B}}(x_{i+1}, y_i) = B_k \leq A_k(f(x^*) - f(x_k)) + \frac{1}{2}\|x^* - x_0\|^2_{\mathbf{B}} \leq \frac{1}{2}\|x^* - x_0\|^2_{\mathbf{B}}. \tag{51}$$

The objective now is to lower bound the quantity $\sum_{i=0}^{k-1}\frac{1}{\hat{r}_{\mathbf{B}}(x_{i+1},y_i)}$ from (50), subject to the constraint given by (51). After defining $\xi_i \overset{\text{def}}{=} \hat{r}_{\mathbf{B}}(x_{i+1}, y_i)$ and $D \overset{\text{def}}{=} \frac{8}{3L_3}\|x_0 - x^*\|^2_{\mathbf{B}}$, our aim is to minimize

$$\xi^* \overset{\text{def}}{=} \min_{\xi\in\mathbb{R}^k}\left\{\sum_{i=0}^{k-1}\frac{1}{\xi_i} \;:\; \sum_{i=0}^{k-1} A_{i+1}\xi_i^4 \leq D\right\}.$$

We may introduce a Lagrange multiplier λ, giving us the following optimality conditions:

$$\frac{1}{\xi_i^2} = \lambda A_{i+1}\xi_i^3, \quad i \in \{0, \ldots, k-1\}.$$

Therefore, $\xi_i = \left(\frac{1}{\lambda A_{i+1}}\right)^{1/5}$. This gives us

$$D = \sum_{i=0}^{k-1} A_{i+1}\left(\frac{1}{\lambda A_{i+1}}\right)^{4/5} = \frac{1}{\lambda^{4/5}}\sum_{i=0}^{k-1} A_{i+1}^{1/5}.$$

Thus, we have

$$\xi^* = \sum_{i=0}^{k-1}(\lambda A_{i+1})^{1/5} = \frac{1}{D^{1/4}}\left(\sum_{i=0}^{k-1} A_{i+1}^{1/5}\right)^{5/4},$$

and so

$$\sum_{i=0}^{k-1} \frac{1}{\hat{r}_{\mathbf{B}}(x_{i+1}, y_i)} \geq \frac{1}{D^{1/4}} \left(\sum_{i=0}^{k-1} A_{i+1}^{1/5} \right)^{5/4}.$$

It follows that

$$A_k \geq \frac{1}{8L_3 D^{1/2}} \left(\sum_{i=1}^{k} A_i^{1/5} \right)^{5/2}, \quad k \geq 1.$$

Let $\theta = \frac{1}{8L_3 D^{1/2}}$ and $C_k = \left(\sum_{i=1}^{k} A_i^{1/5} \right)^{1/2}$. Then, we have that

$$C_{k+1}^2 - C_k^2 \geq \theta^{1/5} C_{k+1}.$$

Thus, we have that $C_1 \geq \theta^{1/5}$, $C_{k+1} \geq C_k$, and so

$$\begin{aligned} \theta^{1/5} C_{k+1} &\leq (C_{k+1} - C_k)(C_{k+1} + C_k) \\ &\leq 2C_{k+1}(C_{k+1} - C_k). \end{aligned}$$

Thus, it follows that $C_k \geq \theta^{1/5}(1 + \frac{1}{2}(k-1))$ for all $k \geq 1$. Taken together, this gives us that

$$A_k \geq \theta \left(C_k^2\right)^{5/2} \geq \theta \left(\theta^{1/5} \frac{k+1}{2} \right)^5 = \theta^2 \left(\frac{k+1}{2} \right)^5 = \frac{3}{256 L_3 \|x_0 - x^*\|_{\mathbf{B}}^2} \left(\frac{k+1}{2} \right)^5.$$

■

Theorem 37 *Suppose there is some $1 \leq i \leq N$ such that for all iterations $1 \leq j < i$, we have that $\rho_{init}^- \leq (1 + \tilde{\varepsilon}_{fs}) \|x_{j+1}^- - y_j^-\|_{\mathbf{B}}^2$ and $\rho_{init}^- \leq \|x_{j+1}^- - y_j^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$, and for iteration i, either*

(a) $\rho_{init}^- > (1 + \tilde{\varepsilon}_{fs}) \|x_{i+1}^- - y_i^-\|_{\mathbf{B}}^2$, *or*

(b) $\rho_{init}^- \leq (1 + \tilde{\varepsilon}_{fs}) \|x_{i+1}^- - y_i^-\|_{\mathbf{B}}^2$ *and* $\rho_{init}^- > \|x_{i+1}^- - y_i^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$.

Then, FastQuartic *returns x_{i+1} such that*

$$f(x_{i+1}) - f(x^*) \leq 2L_3 \rho_{init}^- \|x_0 - x^*\|_{\mathbf{B}}^2. \tag{52}$$

Proof By the algorithm statement, we have that $\tilde{\varepsilon}_{fs} = \min\left\{ \frac{L_3^2 (\rho_{\text{init}}^-)^2}{1000\mathcal{G}}, \frac{1}{2} \right\}$. By $\tilde{\varepsilon}_{aam} > 0$ sufficiently small, we mean that

$$\tilde{\varepsilon}_{aam} \leq \min\left\{ \left(\frac{\tilde{\varepsilon}_{fs}}{\mathcal{V}(1 + \tilde{\varepsilon}_{fs})} \right)^4, \left(\frac{\tilde{\varepsilon}_{fs} \rho_{\text{init}}^-}{\mathcal{Q}(1 + \tilde{\varepsilon}_{fs})} \right)^4, \left(\frac{L_3 (\rho_{\text{init}}^-)^3}{1000\mathcal{W}} \right)^4, \frac{1}{2} \right\}.$$

For both cases (a) and (b), it holds by Lemma 23 (and the statement of this lemma) that

$$A_i f(x_i) + B_i \leq \psi_i^* \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^d} \psi_i(x).$$

We begin by considering the case where (a) holds. We first observe that, since $f(\cdot)$ is convex, we have that, for all $z \in \mathscr{L}$,

$$f(z) - f(x^*) \leq \mathcal{P}^{1/2}\|\nabla f(z)\|_{\mathbf{B}^{-1}}.$$

If $\|\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} < \frac{\varepsilon^2}{\mathcal{P}}$, then we are done, as $f(z) - f(x^*) \leq \varepsilon$, so we consider the case where $\|\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} \geq \frac{\varepsilon^2}{\mathcal{P}}$.

Thus, by Lemma 20, we have that

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \geq \frac{1 - \mathcal{V}\tilde{\varepsilon}_{aam}^{1/4}}{2L_3\hat{r}^2_{\mathbf{B}}(x_{k+1}, y_k)}\|\nabla f(x_{k+1})\|^2_{\mathbf{B}^{-1}} + \frac{3L_3}{8}\hat{r}^4_{\mathbf{B}}(x_{k+1}, y_k),$$

where $\mathcal{V} \overset{\text{def}}{=} \max\limits_{x,y\in\mathscr{L}} \frac{6Z(x,y)W(x,y)\mathcal{P}}{\varepsilon^2 L_3^{1/4}}$.

Since $\rho^-_{\text{init}} > (1+\tilde{\varepsilon}_{fs})\|x^-_{i+1} - y^-_i\|^2_{\mathbf{B}} = (1+\tilde{\varepsilon}_{fs})\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)$ (by (a)), we may follow the same approach as before to arrive at

$$\begin{aligned}
\min_{x\in\mathbb{R}^d}\psi_{i+1}(x) &\geq A_{i+1}f(x_{i+1}) + B_i - \frac{a^2_{i+1}}{2}\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}} + \langle \nabla f(x_{i+1}), A_{i+1}(y_i - x_{i+1})\rangle \\
&\geq A_{i+1}f(x_{i+1}) + B_i - \frac{A_{i+1}}{2L_3\rho^-_{\text{init}}}\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad + A_{i+1}\left(\frac{1 - \mathcal{V}\tilde{\varepsilon}_{aam}^{1/4}}{2L_3\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)}\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}} + \frac{3L_3}{8}\hat{r}^4_{\mathbf{B}}(x_{i+1}, y_i)\right) \\
&> A_{i+1}f(x_{i+1}) + B_i - \frac{A_{i+1}}{2L_3(1+\tilde{\varepsilon}_{fs})\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)}\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}} \\
&\quad + A_{i+1}\left(\frac{1 - \mathcal{V}\tilde{\varepsilon}_{aam}^{1/4}}{2L_3\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)}\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}} + \frac{3L_3}{8}\hat{r}^4_{\mathbf{B}}(x_{i+1}, y_i)\right) \\
&= A_{i+1}f(x_{i+1}) + B_i \\
&\quad + A_{i+1}\left(\frac{\left((1+\tilde{\varepsilon}_{fs})\left(1 - \mathcal{V}\tilde{\varepsilon}_{aam}^{1/4}\right) - 1\right)\|\nabla f(x_{i+1})\|^2_{\mathbf{B}^{-1}}}{2L_3\hat{r}^2_{\mathbf{B}}(x_{i+1}, y_i)} + \frac{3L_3}{8}\hat{r}^4_{\mathbf{B}}(x_{i+1}, y_i)\right).
\end{aligned}$$

Thus, since $\tilde{\varepsilon}_{fs} = \min\left\{\frac{L_3^2(\rho^-_{\text{init}})^2}{1000\mathcal{G}}, \frac{1}{2}\right\}$ and $\tilde{\varepsilon}_{aam} \leq \left(\frac{\tilde{\varepsilon}_{fs}}{\mathcal{V}(1+\tilde{\varepsilon}_{fs})}\right)^4$, it follows that

$$\min_{x\in\mathbb{R}^d}\psi_{i+1}(x) \geq A_{i+1}f(x_{i+1}) + B_i + \frac{3L_3A_{i+1}}{8}\hat{r}^4_{\mathbf{B}}(x_{i+1}, y_i) = A_{i+1}f(x_{i+1}) + B_{i+1}.$$

As before, we may observe that

$$\begin{aligned}
\psi_{i+1}(x) &= \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \sum_{j=0}^{i+1} a_j\left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1}\rangle\right] \\
&\leq \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \sum_{j=0}^{i+1} a_j f(x) \\
&= A_{i+1}f(x) + \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}},
\end{aligned}$$

and so it follows that

$$f(x_{i+1}) - f(x^*) \leq \frac{1}{2A_{i+1}}\|x_0 - x^*\|_{\mathbf{B}}^2.$$

By Lemma 35, we know that $A_{i+1} \geq \frac{1}{4L_3\rho_{\text{init}}^-}$, and so it follows that

$$f(x_{i+1}) - f(x^*) \leq 2L_3\rho_{\text{init}}^-\|x_0 - x^*\|_{\mathbf{B}}^2.$$

We now consider the case where (b) holds, i.e., $\rho_{\text{init}}^- \leq (1+\tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ and $\rho_{\text{init}}^- > \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$. We may observe that

$$\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 \geq \frac{\rho_{\text{init}}^-}{1+\tilde{\varepsilon}_{fs}},$$

and so, if we choose $\tilde{\varepsilon}_{aam} \leq \left(\frac{\tilde{\varepsilon}_{fs}\rho_{\text{init}}^-}{\mathcal{Q}(1+\tilde{\varepsilon}_{fs})}\right)^4$, it follows that

$$\rho_{\text{init}}^- > \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4} \geq \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \frac{\tilde{\varepsilon}_{fs}\rho_{\text{init}}^-}{(1+\tilde{\varepsilon}_{fs})} \geq (1-\tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2,$$

and so we have that

$$(1-\tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 \leq \rho_{\text{init}}^- \leq (1+\tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2.$$

Following a line of reasoning as before, we may use Lemma 22 with $c = (1+\tilde{\varepsilon}_{fs})^{-1}$, along with the fact that $\rho_{\text{init}}^- \geq (1-\tilde{\varepsilon}_{fs})\|x_{i+1}^- - y_i^-\|_{\mathbf{B}}^2$, to see that

$$\begin{aligned}
\min_{x\in\mathbb{R}^d}\psi_{i+1}(x) &\geq A_{i+1}f(x_{i+1}) + B_i - \frac{a_{i+1}^2}{2}\|\nabla f(x_{i+1})\|_{\mathbf{B}^{-1}}^2 + \langle \nabla f(x_{i+1}), A_{i+1}(y_i - x_{i+1})\rangle \\
&\geq A_{i+1}f(x_{i+1}) + B_i - \frac{A_{i+1}}{2L_3\rho_i}\|\nabla f(x_{i+1})\|_{\mathbf{B}^{-1}}^2 \\
&\quad + A_{i+1}\left(\frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{i+1},y_i)}\|\nabla f(x_{i+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{i+1},y_i) - \frac{(1+\tilde{\varepsilon}_{fs})\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{\rho_{\text{init}}^-}\right) \\
&\geq A_{i+1}f(x_{i+1}) + B_i - \frac{A_{i+1}}{2L_3(1-\tilde{\varepsilon}_{fs})\hat{r}_{\mathbf{B}}^2(x_{i+1},y_i)}\|\nabla f(x_{i+1})\|_{\mathbf{B}^{-1}}^2 \\
&\quad + A_{i+1}\left(\frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{i+1},y_i)}\|\nabla f(x_{i+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{i+1},y_i) - \frac{(1+\tilde{\varepsilon}_{fs})\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{\rho_{\text{init}}^-}\right) \\
&= A_{i+1}f(x_{i+1}) + B_i + A_{i+1}\left(\frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{i+1},y_i) - \hat{\varepsilon}_{curr}\right),
\end{aligned}$$

where

$$\hat{\varepsilon}_{curr} \stackrel{\text{def}}{=} \frac{\tilde{\varepsilon}_{fs}}{2L_3(1-\tilde{\varepsilon}_{fs})\rho_{\text{init}}^-}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{(1+\tilde{\varepsilon}_{fs})\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{\rho_{\text{init}}^-}.$$

Thus, for $\tilde{\varepsilon}_{fs} = \min\left\{\frac{L_3^2(\rho_{\text{init}}^-)^2}{1000\mathcal{G}}, \frac{1}{2}\right\}$, and $\tilde{\varepsilon}_{aam} \leq \left(\frac{L_3(\rho_{\text{init}}^-)^3}{1000\mathcal{W}}\right)^4$, it follows that

$$\min_{x\in\mathbb{R}^d}\psi_{i+1}(x) \geq A_{i+1}f(x_{i+1}) + B_i + A_{i+1}\left(\frac{3L_3}{16}\hat{r}_{\mathbf{B}}^4(x_{i+1},y_i)\right) = A_{i+1}f(x_{i+1}) + B_{i+1}.$$

Therefore, it follows that

$$f(x_{i+1}) - f(x^*) \leq \frac{1}{2A_{i+1}}\|x_0 - x^*\|_{\mathbf{B}}^2,$$

and since by Lemma 35, we know that $A_{i+1} \geq \frac{1}{4L_3\rho_{\text{init}}^-}$, we have that

$$f(x_{i+1}) - f(x^*) \leq 2L_3\rho_{\text{init}}^-\|x_0 - x^*\|_{\mathbf{B}}^2.$$

■

Appendix C. Convex quartics and ℓ_4 regression

While we have so far focused on the moderate-accuracy regime, the procedure outlined previously can in fact be used beyond the non-smooth setting to achieve nearly condition number-independent high-accuracy convergence rates for some convex polynomial optimization problems. Specifically, we show how it may be used to solve a large class of convex quartic minimization problems, namely

$$f(x) = c^\top x + x^\top \mathbf{G} x + \mathbf{T}[x,x,x] + \frac{1}{24}\|\mathbf{A}x\|_4^4 \tag{53}$$

for some $c \in \mathbb{R}^d$, $\mathbf{G} \in \mathbb{R}^{d\times d}$, $\mathbf{T} \in \mathbb{R}^{d\times d\times d}$, and $\mathbf{A} \in \mathbb{R}^{n\times d}$ such that $\mathbf{A}^\top\mathbf{A} \succ 0$. We call these functions *structured convex quartics*. Notably, this class includes the problems of ℓ_4 regression, which is in turn an instance of the more general problem of ℓ_p regression Dasgupta et al. (2009); Cohen and Peng (2015); Bubeck et al. (2018a); Adil et al. (2019a,b). In the general case, it is known to be NP-hard to find the global minimizer of a quartic polynomial (Murty and Kabadi, 1987; Parrilo and Sturmfels, 2003), or even to decide if the quartic polynomial is convex (Ahmadi et al., 2013), but here we assume that $f(\cdot)$ is convex.

In order to get a handle on the regularity properties of $f(\cdot)$, we establish its third-order smoothness and fourth-order uniform convexity parameters w.r.t. $\|\cdot\|_{\mathbf{A}^\top\mathbf{A}}$.

Lemma 38 (Third-order smoothness) *Suppose $f(\cdot)$ is of the form (53). Then, for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla^3 f(y) - \nabla^3 f(x)\|_{\mathbf{A}^\top\mathbf{A}}^* \leq \|y - x\|_{\mathbf{A}^\top\mathbf{A}}. \tag{54}$$

Proof Note that for all $\xi \in \mathbb{R}^d$,

$$\|\nabla^4 f(\xi)\|_{\mathbf{B}}^* = \max_{h:\|h\|_{\mathbf{B}}\leq 1}\left|\nabla^4 f(\xi)[h]^4\right| = \max_{h:\|h\|_{\mathbf{B}}\leq 1}\|\mathbf{A}h\|_4^4 \leq \max_{h:\|h\|_{\mathbf{B}}\leq 1}\|\mathbf{A}h\|_2^4. \tag{55}$$

Setting $\mathbf{B} = \mathbf{A}^\top\mathbf{A}$ gives us

$$\max_{h:\|h\|_{\mathbf{A}^\top\mathbf{A}}\leq 1}\|\mathbf{A}h\|_2^4 \leq 1.$$

By the mean value theorem, we have, for some ξ along the line between x and y,

$$\frac{\|\nabla^3 f(y) - \nabla^3 f(x)\|_{\mathbf{A}^\top\mathbf{A}}^*}{\|y - x\|_{\mathbf{A}^\top\mathbf{A}}} = \|\nabla^4 f(\xi)\|_{\mathbf{A}^\top\mathbf{A}}^* \leq 1,$$

and so it follows that

$$\|\nabla^3 f(y) - \nabla^3 f(x)\|_{\mathbf{A}^\top\mathbf{A}}^* \leq \|y - x\|_{\mathbf{A}^\top\mathbf{A}}.$$

■

Lemma 39 (Order 4 uniform convexity) *Suppose $f(\cdot)$ is of the form (53). Then, for all $x, y \in \mathbb{R}^d$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{72n} \|y - x\|^4_{\mathbf{A}^\top \mathbf{A}}. \tag{56}$$

Proof Following the same idea as in the proof of Lemma 38, we note that, for all $x, y \in \mathbb{R}^d$,

$$f(y) = \Phi_{x,4}(y).$$

Since $f(\cdot)$ is convex by definition, it follows that

$$0 \preceq \nabla^2 f(y) = \nabla^2 f(x) + \nabla^3 f(x)[y - x] + \frac{1}{2}\nabla^4 f(x)[y - x, y - x].$$

Let $h = y - x$. Then, following the approach of Nesterov (2018a), we have

$$-\nabla^3 f(x)[h] \preceq \nabla^2 f(x) + \frac{1}{2}\nabla^4 f(x)[h, h].$$

Since this holds for any x, y (and therefore, for any direction h), we can replace h with τh for any $\tau > 0$ and arrive at

$$-\tau\nabla^3 f(x)[h] \preceq \nabla^2 f(x) + \tau^2\frac{1}{2}\nabla^4 f(x)[h, h].$$

Furthermore, we can replace h by $-h$ to get

$$\tau\nabla^3 f(x)[h] \preceq \nabla^2 f(x) + \tau^2\frac{1}{2}\nabla^4 f(x)[h, h],$$

and so after dividing through by τ, we obtain

$$-\frac{1}{\tau}\nabla^2 f(x) - \frac{\tau}{2}\nabla^4 f(x)[h, h] \preceq \nabla^3 f(x)[h] \preceq \frac{1}{\tau}\nabla^2 f(x) + \frac{\tau}{2}\nabla^4 f(x)[h, h].$$

We may now observe that

$$\begin{aligned}
f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\nabla^2 f(x)[y - x, y - x] + \frac{1}{6}\nabla^3 f(x)[y - x]^3 \\
&\quad + \frac{1}{24}\nabla^4 f(x)[y - x]^4 \\
&\geq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{1}{2} - \frac{1}{6\tau}\right)\nabla^2 f(x)[y - x, y - x] \\
&\quad + \left(\frac{1}{24} - \frac{\tau}{12}\right)\nabla^4 f(x)[y - x]^4.
\end{aligned}$$

Setting $\tau = \frac{1}{3}$ gives us

$$\begin{aligned}
f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{72}\nabla^4 f(x)[y - x]^4 \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{72}\|\mathbf{A}(y - x)\|_4^4 \\
&\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{72n}\|\mathbf{A}(y - x)\|_2^4 \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{72n}\|y - x\|^4_{\mathbf{A}^\top \mathbf{A}},
\end{aligned}$$

where the second inequality follows from the fact that, for $v \in \mathbb{R}^n$, $\|v\|_2 \leq n^{\frac{1}{4}}\|v\|_4$. Thus, we arrive at (56). ■

Combining Theorem 26 with the appropriate notion of uniform convexity, we may establish a rate of linear convergence, based on the (fourth-order) condition number $\kappa_4 \stackrel{\text{def}}{=} \frac{L_3}{\mu_4}$, by relying on an additional restarting procedure (Algorithm 2). With this result in hand, the proof of the main theorem follows almost immediately.

Theorem 40 *Suppose $f(x)$ is third-order L_3-smooth and fourth-order μ_4-uniformly convex w.r.t. $\|\cdot\|_{\mathbf{B}}$. Then, under appropriate initialization,* FastQuartic + Restarting *(Algorithm 2) finds a point x_N such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

in $O\left(\kappa_4^{1/5} \log\left(\frac{f(x_0)-f(x^*)}{\varepsilon}\right)\right)$ iterations, where each iteration requires $O(\log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver, and where $\mathcal{Z}$ is a polynomial in various problem-dependent parameters.

Proof Begin by running the FastQuartic algorithm for $N_{inner} = \left\lceil \left(\frac{512L_3}{\mu_4}\right)^{1/5} \right\rceil$ iterations, as in each (outer) iteration of Algorithm 2. By combining Theorem 26 with the fact that $f(\cdot)$ is uniformly convex, we have that, for any $k \geq 0$,

$$f(x_{k+1}) - f(x^*) \leq \frac{128L_3\|x_k - x^*\|_{\mathbf{B}}^4}{3}\left(\frac{2}{N_{inner}+1}\right)^5 \leq \frac{512L_3(f(x_k)-f(x^*))}{3\mu_4(N_{inner})^5}.$$

It follows from our choice of N_{inner} that

$$f(x_{k+1}) - f(x^*) \leq \frac{f(x_k)-f(x^*)}{2}.$$

Because we reduce the optimality gap by a constant factor each iteration of Algorithm 2, it suffices to run FastQuartic + Restarting for $N = O\left(\log\left(\frac{f(x_0)-f(x^*)}{\varepsilon}\right)\right)$ iterations to achieve a point x_N such that

$$f(x_N) - f(x^*) \leq \varepsilon,$$

which gives a total iteration complexity of $O(N_{inner} \cdot N) = O\left(\kappa_4^{1/5} \log\left(\frac{f(x_0)-f(x^*)}{\varepsilon}\right)\right)$. ■

Having developed all of the necessary results, we may now prove our main theorem for structured convex quartics, as well as the natural corollary regarding the special case of ℓ_4 regression.

Theorem 41 *Let $f(\cdot)$ be a convex function of the form (53). Then, under appropriate initialization,* FastQuartic *finds a point x_N such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

with total computational cost $O(n^{1/5}(\text{GO}+\text{LSS})\log^{O(1)}(\mathcal{Z}/\varepsilon))$, where GO *is the time to calculate the gradient of $f(\cdot)$,* LSS *is the time to solve a $d \times d$ linear system, and $\mathcal{Z}$ is a polynomial in various problem-dependent parameters.*

Algorithm 2 FastQuartic + Restarting

Input: $\varepsilon > 0$, $x_0 = 0$, $A_0 = 0$, $\mathbf{B} \succ 0$, $\frac{1}{2} > \rho_{\text{init}}^{-} > 0$, $\rho_{\text{init}}^{+} = \mathcal{P}$, $\tilde{\varepsilon}_{aam} > 0$, N, $N_{inner} = O\left((L_3/\mu_4)^{1/5}\right)$.
for $k = 0$ **to** N **do**
  $x_{k+1} = \mathsf{FastQuartic}(\varepsilon, x_k, A_0, \mathbf{B}, \rho_{\text{init}}^{-}, \rho_{\text{init}}^{+}, \tilde{\varepsilon}_{aam}, N_{inner})$
end for
return x_{N+1}

Proof [Proof of Theorem 41] The proof follows by combining Theorem 40 with Lemmas 38 and 39. ■

Corollary 42 *For the problem of ℓ_4 regression, i.e., problems of the form*

$$\min_{x \in \mathbb{R}^d} f(x) = c^\top x + \|\mathbf{A}x - b\|_4^4,$$

for $c \in \mathbb{R}^d$, $b \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times d}$ such that $\mathbf{A}^\top \mathbf{A} \succ 0$, FastQuartic *finds, under appropriate initialization, a point x_N such that*

$$f(x_N) - f(x^*) \leq \varepsilon$$

with $O(n^{1/5} \log^{O(1)}(\mathcal{Z}/\varepsilon))$ calls to a gradient oracle and linear system solver.

Proof [Proof of Corollary 42] Note that for all $x \in \mathbb{R}^d$, $\nabla^4 f(x) = 24 \sum_{i=1}^{n} a_i^{\otimes 4}$, where $\mathbf{A} = [a_1 a_2 \dots a_n]^\top$. Since $f(x)$ is a convex quartic function, we may equivalently express it as its fourth-order Taylor expansion

$$\begin{aligned} f(x) &= f(0) + \nabla f(0)^\top x + \frac{1}{2} x^\top \nabla^2 f(0) x + \frac{1}{6} \nabla^3 f(0)[x,x,x] + \frac{1}{24} \nabla^4 f(0)[x]^{\otimes 4} \\ &= f(0) + \nabla f(0)^\top x + \frac{1}{2} x^\top \nabla^2 f(0) x + \frac{1}{6} \nabla^3 f(0)[x,x,x] + \|\mathbf{A}x\|_4^4, \end{aligned}$$

and so since $f(\cdot)$ is of the form (53), for $\mathbf{A}^\top \mathbf{A} \succ 0$, the result follows from Theorem 41. ■

Appendix D. Proofs for Section 3

D.1. Proof of Lemmas 2 and 3

Proof [Proof of Lemma 2] Since $|c| = \max\{c, -c\}$, it follows by Fact (1) that $|c| \leq \text{sabs}_\mu(c) \leq |c| + \mu$. Thus, we have that

$$\|x\|_1 = \sum_{i=1}^{m} |x_i| \leq \sum_{i=1}^{m} \text{sabs}_\mu(x_i) \leq \|x\|_1 + \mu m. \tag{57}$$

■

Proof [Proof of Lemma 3] The proof follows immediately from Fact (1). ■

D.2. Proof of Theorem 5

Proof [Proof of Theorem 5] Recall that $\operatorname{smax}_\mu(x) = \mu f(Z_\mu(x))$ for $f(z) = \mu \log(z)$. First, we establish some preliminary observations concerning the higher-order derivatives of $Z_\mu(x)$ and $f(z)$. To begin, note that for $k \geq 1$,

$$f^{(k)}(z) = \frac{(-1)^{k-1}(k-1)!\mu}{z^k}. \tag{58}$$

In addition, since $Z_\mu(x)$ is a separable function, it follows that, for $k \geq 1$,

$$\left[\nabla^k Z_\mu(x)\right]_{i_1,\ldots,i_k} = \begin{cases} \frac{e^{\frac{x_i}{\mu}}}{\mu^k}, & \text{if } i_1 = i_2 = \cdots = i_k = i, \quad \text{for } i \in [m]\,; \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

Now, letting Π_p denote the set of all partitions on p elements, we may observe that

$$\begin{aligned}
\nabla^p \operatorname{smax}_\mu(x)[h]^p &= \nabla^p f(Z_\mu(x))[h]^p \\
&= \sum_{\pi \in \Pi_p} f^{|\pi|}(Z_\mu(x)) \cdot \prod_{B \in \pi} \nabla^{|B|} Z_\mu(x)[h]^{|B|} \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \nabla^{|B|} Z_\mu(x)[h]^{|B|} \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \sum_{i_1,i_2,\ldots,i_{|B|}=1}^{m} \left(\left[\nabla^{|B|} Z_\mu(x)\right]_{i_1,i_2,\ldots,i_{|B|}} \prod_{j=1}^{|B|} h_{i_j} \right) \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \sum_{i=1}^{m} \left(\left[\nabla^{|B|} Z_\mu(x)\right]_{i,i,\ldots,i} h_i^{|B|} \right) \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \sum_{i=1}^{m} \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right),
\end{aligned}$$

where the second equality follows from Faà di Bruno's formula, the third equality follows from (58), and the final two equalities follow from (59). For convenience, we denote

$$h^{|B|} \stackrel{\text{def}}{=} \left[h_1^{|B|},\ h_2^{|B|}, \ldots,\ h_m^{|B|}\right]^\top.$$

As our goal is to bound $|\nabla^p \operatorname{smax}_\mu(x)[h]^p|$, we may observe that, by the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned}
|\nabla^p \operatorname{smax}_\mu(x)[h]^p| &= \left| \mu \sum_{\pi \in \Pi_p} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \sum_{i=1}^p \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right) \right| \\
&\leq \mu \sum_{\pi \in \Pi_p} \left| \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \sum_{i=1}^p \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right) \right| \\
&\leq \mu \sum_{\pi \in \Pi_p} \left| \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \right| \cdot \prod_{B \in \pi} \left| \sum_{i=1}^p \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right) \right| \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \left| \sum_{i=1}^p \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right) \right|.
\end{aligned}$$

Finally, using Hölder's inequality and simplifying, we have that

$$\begin{aligned}
|\nabla^p \operatorname{smax}_\mu(x)[h]^p| &= \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \left| \sum_{i=1}^p \left(\frac{e^{\frac{x_i}{\mu}}}{\mu^{|B|}} h_i^{|B|} \right) \right| \\
&\leq \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \left(\frac{Z_\mu(x)}{\mu^{|B|}} \|h^{|B|}\|_\infty \right) \\
&\leq \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \prod_{B \in \pi} \left(\frac{Z_\mu(x)}{\mu^{|B|}} \|h\|_2^{|B|} \right) \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)!}{(Z_\mu(x))^{|\pi|}} \cdot \left(\frac{(Z_\mu(x))^{|\pi|} \|h\|_2^p}{\mu^p} \right) \\
&= \mu \sum_{\pi \in \Pi_p} \frac{(|\pi|-1)! \|h\|_2^p}{\mu^p} \\
&\leq \frac{(p-1)! \|h\|_2^p}{\mu^{p-1}} \sum_{\pi \in \Pi_p} 1 \\
&= \frac{B_p (p-1)! \|h\|_2^p}{\mu^{p-1}},
\end{aligned}$$

where $B_p \overset{\text{def}}{=} |\Pi_p|$ is the p^{th} Bell number, i.e., the number of partitions on p elements. Since $B_p \leq \left(\frac{p}{\ln(p+1)} \right)^p$ (Berend and Tassa, 2010), it follows that

$$|\nabla^p \operatorname{smax}_\mu(x)[h]^p| \leq \frac{\left(\frac{p}{\ln(p+1)} \right)^p (p-1)! \|h\|_2^p}{\mu^{p-1}}. \tag{60}$$

■

D.3. Proof of Lemma 6

Proof [Proof of Lemma 6] Note that, for all $\zeta \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla^{p+1} f(\zeta)\|^*_{\mathbf{A}^\top\mathbf{A}} &= \max_{h:\|h\|_{\mathbf{A}^\top\mathbf{A}}\le 1} \left|\nabla^{p+1} f(\zeta)[h]^{p+1}\right| \\ &= \max_{h:\|\mathbf{A}h\|_2\le 1} \left|\nabla^{p+1} f(\zeta)[h]^{p+1}\right| \\ &\le \max_{h:\|\mathbf{A}h\|_2\le 1} L_p\|\mathbf{A}h\|_2^{p+1} \\ &= L_p,\end{aligned}$$

where the inequality follows from (16). Thus, we may see by a standard mean value theorem argument that, for any $x, y \in \mathbb{R}^d$,

$$\|\nabla^p f(y) - \nabla^p f(x)\|^*_{\mathbf{A}^\top\mathbf{A}} \le L_p\|y - x\|_{\mathbf{A}^\top\mathbf{A}}.$$

■

D.4. Proof of Theorem 7

Proof [Proof of Theorem 7] By applying the chain rule to $f(\cdot)$ p times, we may observe that

$$\nabla^p f(x)[h]^p = \nabla^p \operatorname{smax}_\mu(\mathbf{A}x - b)[\mathbf{A}h]^p, \tag{61}$$

and so it follows from (2) and Lemma 16 that $f(x)$ is (order p) $\frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!}{\mu^p}$-smooth w.r.t. $\|\cdot\|_{\mathbf{A}^\top\mathbf{A}}$. ■

D.5. Proof of Theorem 8

Proof [Proof of Theorem 8] First, we observe that since soft-$\ell 1_\mu(x) = \sum_{i=1}^d \operatorname{sabs}_\mu(x_i)$, it follows by Theorem 5 that

$$\left|\nabla^{p+1} \text{soft-}\ell 1_\mu(x)[h]^{p+1}\right| \le \sum_{i=1}^d \left|\nabla^{p+1} \operatorname{sabs}_\mu(x_i)[h]^{p+1}\right| \le \frac{d\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!}{\mu^p}\|h\|_2^{p+1}. \tag{62}$$

In addition, we may note that

$$\begin{aligned}\left|\nabla^{p+1} \operatorname{softSVM}_\mu(\tilde{\mathbf{Q}}x)[\tilde{\mathbf{Q}}h]^{p+1}\right| &\le \frac{1}{m}\sum_{i=1}^m \frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!\|\tilde{\mathbf{Q}}h\|_2^{p+1}}{\mu^p} \\ &\le \frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p!\|\tilde{\mathbf{Q}}^\top\tilde{\mathbf{Q}}\|^{\frac{p+1}{2}}}{\mu^p}\|h\|_2^{p+1}.\end{aligned}$$

Taken together, we may see that

$$\begin{aligned}\left|\nabla^{p+1} f_\mu(x)[h]^{p+1}\right| &= \left|\lambda \nabla^{p+1} \text{soft-}\ell 1_\mu(x)[h]^{p+1} + \nabla^{p+1} \text{softSVM}_\mu(\tilde{\mathbf{Q}}x)[\tilde{\mathbf{Q}}h]^{p+1}\right| \\ &\le \lambda \left|\nabla^{p+1} \text{soft-}\ell 1_\mu(x)[h]^{p+1}\right| + \left|\nabla^{p+1} \text{softSVM}_\mu(\tilde{\mathbf{Q}}x)[\tilde{\mathbf{Q}}h]^{p+1}\right| \\ &\le \frac{\left(\frac{p+1}{\ln(p+2)}\right)^{p+1} p! \left(\lambda d + \|\tilde{\mathbf{Q}}^\top \tilde{\mathbf{Q}}\|^{\frac{p+1}{2}}\right)}{\mu^p} \|h\|_2^{p+1},\end{aligned}$$

and so the theorem follows from Lemma 6. ■

Appendix E. Section 4 Algorithms

Algorithm 3 ApproxAuxMin

Input: y_k, $\tilde{\varepsilon}_{aam} > 0$, $K = O(\log(\mathcal{A}/\tilde{\varepsilon}_{aam}))$, $h_0 = 0$.
for $t = 0$ **to** K **do**
  $c_t \overset{\text{def}}{=} \nabla f(y_k) + \nabla^2 f(y_k) h_t + \frac{1}{2}\nabla^3 f(y_k)[h_t]^2 + L_3 \|h_t\|_{\mathbf{B}}^2 \mathbf{B} h_t$
  $h_{t+1} = \text{argmin}_{h \in \mathbb{R}^d} \left\{ \langle c_t, h - h_t \rangle + \frac{1}{\sqrt{2}} (h - h_t)^\top \nabla^2 f(y_k)(h - h_t) + \frac{\sqrt{2} L_3}{4} \|h - h_t\|_{\mathbf{B}}^4 \right\}$
end for
return $x_{k+1} = y_k + h_K$

Algorithm 4 RhoSearch

Input: x_k, v_k, A_k, ρ_{init}^+, ρ_{init}^- (s.t. $\rho_{\text{init}}^+ \ge \rho_k^* \ge \rho_{\text{init}}^-$), $\tilde{\varepsilon}_{rs} > 0$, $\tilde{\varepsilon}_{aam} > 0$, $M = O(\log(\mathcal{R}/\tilde{\varepsilon}_{rs}))$.
Define $\tilde{\delta} \overset{\text{def}}{=} 6 \left(\frac{\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4} \mathcal{P}^{1/2} + \left(\frac{12 \tilde{\varepsilon}_{aam}}{L_3}\right)^{1/2}$.
$\rho^+ \leftarrow \rho_{\text{init}}^+$, $\rho^- \leftarrow \rho_{\text{init}}^-$
for $t = 1$ **to** M **do**
  $\hat{\rho} = \frac{\rho^- + \rho^+}{2}$
  $\hat{a}_{k+1} = \frac{1 + \sqrt{1 + 4 L_3 A_k \hat{\rho}}}{2 L_3 \hat{\rho}}$ $\quad \left(\implies \hat{a}_{k+1}^2 = \frac{A_k + \hat{a}_{k+1}}{L_3 \hat{\rho}} \right)$
  $A_{k+1} = A_k + \hat{a}_{k+1}$
  $\tau_k = \frac{\hat{a}_{k+1}}{A_{k+1}}$
  $\hat{y}_k = (1 - \tau_k) x_k + \tau_k v_k$
  $\hat{x}_{k+1} \leftarrow \text{ApproxAuxMin}(\hat{y}_k, \tilde{\varepsilon}_{aam})$
  if $\hat{\rho} > \hat{\zeta}(\hat{\rho}) + \tilde{\delta}$ **then**
    $\rho^+ \leftarrow \hat{\rho}$
  else if $\hat{\rho} < \hat{\zeta}(\hat{\rho}) - \tilde{\delta}$ **then**
    $\rho^- \leftarrow \hat{\rho}$
  else
    return $\hat{\rho}, \hat{x}_{k+1}, \hat{a}_{k+1}$
  end if
end for
return $\rho^-, \hat{x}_{k+1}, \hat{a}_{k+1}$

Algorithm 5 FastQuartic

Input: $\varepsilon > 0$, $x_0 = 0$, $A_0 = 0$, $\mathbf{B} \succ 0$, $\frac{1}{2} > \rho_{\text{init}}^- > 0$, $\rho_{\text{init}}^+ = \mathcal{P}$, $\tilde{\varepsilon}_{aam} > 0$, N.
Define $\psi_0(x) \overset{\text{def}}{=} \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2$, $\tilde{\varepsilon}_{fs} \overset{\text{def}}{=} \min\left\{\frac{L_3^2(\rho_{\text{init}}^-)^2}{1000\mathcal{G}}, \frac{1}{2}\right\}$, $\tilde{\varepsilon}_{rs} \overset{\text{def}}{=} \min\left\{\frac{L_3(\rho_{\text{init}}^-)^3}{1000\mathcal{T}}, \rho_{\text{init}}^-, \frac{1}{2}\right\}$, $\mathcal{T}$ as in (66).
for $k = 0$ **to** N **do**
  $v_k = \operatorname{argmin}_{x \in \mathbb{R}^d} \psi_k(x)$
  $a_{k+1}^- = \frac{1+\sqrt{1+4L_3A_k\rho_{\text{init}}^-}}{2L_3\rho_{\text{init}}^-}$ $\qquad \left(\implies (a_{k+1}^-)^2 = \frac{A_k + a_{k+1}^-}{L_3\rho_{\text{init}}^-}\right)$
  $A_{k+1}^- = A_k + a_{k+1}^-$
  $\tau_k^- = \frac{a_{k+1}^-}{A_{k+1}^-}$
  $y_k^- = (1 - \tau_k^-)x_k + \tau_k^- v_k$
  $x_{k+1}^- \leftarrow \mathsf{ApproxAuxMin}(y_k^-, \tilde{\varepsilon}_{aam})$
  if $\rho_{\text{init}}^- > (1 + \tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ **then**
    return x_{k+1}^-
  else if $\rho_{\text{init}}^- \le (1 + \tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ and $\rho_{\text{init}}^- > \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$ ($\mathcal{Q}$ as defined in (45)) **then**
    return x_{k+1}^-
  else
    $\rho_k, x_{k+1}, a_{k+1} \leftarrow \mathsf{RhoSearch}(x_k, v_k, A_k, \rho_{\text{init}}^+, \rho_{\text{init}}^-, \tilde{\varepsilon}_{rs}, \tilde{\varepsilon}_{aam})$
    $\psi_{k+1} = \psi_k + a_{k+1}\left[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle\right]$
  end if
end for
return x_{N+1}

Appendix F. Proofs for Section 4

F.1. Proof of Lemma 15

Proof To begin, we define

$$Z(x,y) \overset{\text{def}}{=} \|\nabla f(x) + L_3\hat{r}_{\mathbf{B}}^2(x,y)\mathbf{B}(x-y)\|_{\mathbf{B}^{-1}},$$

$$W(x,y) \overset{\text{def}}{=} \left(\|\mathbf{B}^{-1/2}\|^2\|\mathbf{H}(x,y)\|\|\mathbf{B}^{-1}\| + L_3\|x - T_{\mathbf{B}}(y)\|_{\mathbf{B}}^2\right),$$

and

$$\mathbf{H}(x,y) \overset{\text{def}}{=} \nabla^2\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y)) + \frac{1}{2}\nabla^3\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x - T_{\mathbf{B}}(y)].$$

Let $x, y \in \mathbb{R}^d$, let $\hat{r}_{\mathbf{B}}(x,y) \overset{\text{def}}{=} \|x - y\|_{\mathbf{B}}$, and let $\delta(x,y) \overset{\text{def}}{=} \nabla\Omega_{y,\mathbf{B}}(x)$. Using the third-order L_3-smoothness of $f(x)$, we have by (26) and the triangle inequality that

$$\begin{aligned}\left|\|\nabla f(x) + L_3\hat{r}_{\mathbf{B}}^2(x,y)\mathbf{B}(x-y)\|_{\mathbf{B}^{-1}} - \|\delta(x,y)\|_{\mathbf{B}^{-1}}\right| &\le \|\nabla f(x) + L_3\hat{r}_{\mathbf{B}}^2(x,y)\mathbf{B}(x-y) - \delta(x,y)\|_{\mathbf{B}^{-1}} \\ &= \|\nabla f(x) - \nabla\Phi_y(x)\|_{\mathbf{B}^{-1}} \\ &\le \frac{L_3}{6}\hat{r}_{\mathbf{B}}^3(x,y),\end{aligned}$$

where the last inequality follows from (26). Squaring both sides gives us

$$\|\nabla f(x) + L_3 \hat{r}_{\mathbf{B}}^2(x,y)\mathbf{B}(x-y)\|_{\mathbf{B}^{-1}}^2 - \Delta(x,y) \leq \frac{L_3^2}{36}\hat{r}_{\mathbf{B}}^6(x,y),$$

where

$$\Delta(x,y) \stackrel{\text{def}}{=} 2Z(x,y)\|\delta(x,y)\|_{\mathbf{B}^{-1}} - \|\delta(x,y)\|_{\mathbf{B}^{-1}}^2$$

and

$$Z(x,y) \stackrel{\text{def}}{=} \|\nabla f(x) + L_3 \hat{r}_{\mathbf{B}}^2(x,y)\mathbf{B}(x-y)\|_{\mathbf{B}^{-1}}.$$

After expanding and rearranging the terms in the inequality, we arrive at

$$\|\nabla f(x)\|_{\mathbf{B}^{-1}}^2 + \frac{35}{36}L_3^2\hat{r}_{\mathbf{B}}^6(x,y) - \Delta(x,y) \leq 2L_3\hat{r}_{\mathbf{B}}^2(x,y)\langle \nabla f(x), y-x\rangle.$$

Dividing both sides by $2L_3\hat{r}_{\mathbf{B}}^2(x,y)$ gives us

$$\frac{\|\nabla f(x)\|_{\mathbf{B}^{-1}}^2}{2L_3\hat{r}_{\mathbf{B}}^2(x,y)} + \frac{35}{72}L_3\hat{r}_{\mathbf{B}}(x,y)^4 - \frac{\Delta(x,y)}{2L_3\hat{r}_{\mathbf{B}}^2(x,y)} \leq \langle \nabla f(x), y-x\rangle. \tag{63}$$

All that remains is to bound $\Delta(x,y)$. Note that, by (26) and using the fact that $\nabla\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y)) = 0$,

$$\begin{aligned}
&\|\nabla\Omega_{y,\mathbf{B}}(x) - \nabla\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y)) - \nabla^2\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x-T_{\mathbf{B}}(y)] - \frac{1}{2}\nabla^3\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x-T_{\mathbf{B}}(y)]^2\|_{\mathbf{B}^{-1}} \\
&= \|\nabla\Omega_{y,\mathbf{B}}(x) - \nabla^2\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x-T_{\mathbf{B}}(y)] - \frac{1}{2}\nabla^3\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x-T_{\mathbf{B}}(y)]^2\|_{\mathbf{B}^{-1}} \\
&\leq L_3\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}}^3.
\end{aligned}$$

By triangle inequality and rearranging, we have

$$\|\nabla\Omega_{y,\mathbf{B}}(x)\|_{\mathbf{B}^{-1}} \leq \|\mathbf{H}(x,y)(x-T_{\mathbf{B}}(y))\|_{\mathbf{B}^{-1}} + L_3\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}}^3 \tag{64}$$

where $\mathbf{H}(x,y) \stackrel{\text{def}}{=} \nabla^2\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y)) + \frac{1}{2}\nabla^3\Omega_{y,\mathbf{B}}(T_{\mathbf{B}}(y))[x-T_{\mathbf{B}}(y)]$. Note that, by our choice of $\mathbf{B}$, we may write its eigendecomposition as $\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}^\top$, and we may define $\mathbf{B}^{1/2} \stackrel{\text{def}}{=} \mathbf{U}\Lambda^{1/2}\mathbf{U}^\top$ and $\mathbf{B}^{-1/2} \stackrel{\text{def}}{=} \mathbf{U}\Lambda^{-1/2}\mathbf{U}^\top$. Thus, we can then rewrite

$$\begin{aligned}
\|\mathbf{H}(x,y)(x-T_{\mathbf{B}}(y))\|_{\mathbf{B}^{-1}} &= \|\mathbf{B}^{-1/2}\mathbf{H}(x,y)(x-T_{\mathbf{B}}(y))\| \\
&\leq \|\mathbf{B}^{-1/2}\|\|\mathbf{H}(x,y)\|\|x-T_{\mathbf{B}}(y)\| \\
&= \|\mathbf{B}^{-1/2}\|\|\mathbf{H}(x,y)\|\|\mathbf{B}^{-1}\|\|\mathbf{B}^{-1/2}\mathbf{B}^{1/2}(x-T_{\mathbf{B}}(y))\| \\
&\leq \|\mathbf{B}^{-1/2}\|\|\mathbf{H}(x,y)\|\|\mathbf{B}^{-1}\|\|\mathbf{B}^{-1/2}\|\|\mathbf{B}^{1/2}(x-T_{\mathbf{B}}(y))\| \\
&= \|\mathbf{B}^{-1/2}\|^2\|\mathbf{H}(x,y)\|\|\mathbf{B}^{-1}\|\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}},
\end{aligned}$$

and so it follows that

$$\begin{aligned}
\|\nabla\Omega_{y,\mathbf{B}}(x)\|_{\mathbf{B}^{-1}} &\leq \left(\|\mathbf{B}^{-1/2}\|^2\|\mathbf{H}(x,y)\|\|\mathbf{B}^{-1}\| + L_3\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}}^2\right)\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}} \\
&= W(x,y)\|x-T_{\mathbf{B}}(y)\|_{\mathbf{B}}.
\end{aligned}$$

Taken together with (63), we have that

$$\langle \nabla f(x), y - x \rangle \geq \frac{\|\nabla f(x)\|^2_{\mathbf{B}^{-1}}}{2L_3 \hat{r}^2_{\mathbf{B}}(y)} + \frac{35}{72} L_3 \hat{r}_{\mathbf{B}}(y)^4 - \frac{2Z(x,y)W(x,y)\|x - T_{\mathbf{B}}(y)\|_{\mathbf{B}}}{2L_3 \hat{r}^2_{\mathbf{B}}(y)} .$$

■

F.2. Proof of Lemma 16.

Proof We note that, for all $y, z \in \mathbb{R}^d$, since $\Omega_{x,\mathbf{B}}(z)$ is a convex quartic, it similarly follows from the proof of Lemma 39 that

$$\begin{aligned}
\Omega_{x,\mathbf{B}}(z) &= \Omega_{x,\mathbf{B}}(y) + \langle \nabla \Omega_{x,\mathbf{B}}(y), z - y \rangle + \frac{1}{2}(z-y)^\top \nabla^2 \Omega_{x,\mathbf{B}}(y)(z-y) + \frac{1}{6}\nabla^3 \Omega_{x,\mathbf{B}}(y)[z-y]^3 \\
&\quad + \frac{1}{24}\nabla^4 \Omega_{x,\mathbf{B}}(y)[z-y]^4 \\
&\geq \Omega_{x,\mathbf{B}}(y) + \langle \nabla \Omega_{x,\mathbf{B}}(y), z - y \rangle + \frac{1}{72}\nabla^4 \Omega_{x,\mathbf{B}}(y)[z-y]^4 \\
&= \Omega_{x,\mathbf{B}}(y) + \langle \nabla \Omega_{x,\mathbf{B}}(y), z - y \rangle + \frac{L_3}{12}\|z-y\|^4_{\mathbf{B}}.
\end{aligned}$$

■

F.3. Proof of Corollary 18.

Proof We first note that $T_{\mathbf{B}}(y_k) = y_k + h^*$, and so $\Omega_{y_k,\mathbf{B}}(x_{k+1}) - \Omega_{y_k,\mathbf{B}}(T_{\mathbf{B}}(y_k)) = \Gamma_{y_k,\mathbf{B}}(h_t) - \Gamma_{y_k,\mathbf{B}}(h^*)$. As observed by Nesterov (2018a) (see also: Appendix A in (Agarwal et al., 2017)), c_t can be calculated in time proportional to the cost of evaluating $f(\cdot)$. In addition, Nesterov (2018a) notes that (30) can be found by any reasonable linearly convergent procedure, and so given access to the gradient of $w(\lambda)$, this problem can be optimized (to sufficiently small error) in $O(\log^{O(1)}(1/\tilde{\varepsilon}_{aam}))$ calls to a gradient oracle. Since

$$\frac{d}{d\lambda} w(\lambda) = \lambda - \frac{\sqrt{2}}{2} c_t^\top (\sqrt{2}\lambda \mathbf{B} + \nabla^2 f(x))^{-1} \mathbf{B} (\sqrt{2}\lambda \mathbf{B} + \nabla^2 f(x))^{-1} c_t,$$

calculating the gradient requires $O(\text{LSS})$ time.

Finally, since $K = O(\log(\mathcal{A}/\tilde{\varepsilon}_{aam}))$, by our choice of $\mathcal{A}$, it follows from Theorem 17 that

$$\Omega_{y_k,\mathbf{B}}(x_{k+1}) - \Omega_{y_k,\mathbf{B}}(T_{\mathbf{B}}(y_k)) \leq \tilde{\varepsilon}_{aam}.$$

■

F.4. Proof of Lemma 19.

Proof By Lemma 16, we know that

$$\begin{aligned}\Omega_{y_k,\mathbf{B}}(x_{k+1}) - \Omega_{y_k,\mathbf{B}}(T_{\mathbf{B}}(y_k)) &\geq \langle \nabla\Omega_{y_k,\mathbf{B}}(T_{\mathbf{B}}(y_k)), x_{k+1} - T_{\mathbf{B}}(y_k)\rangle + \frac{L_3}{12}\|x_{k+1} - T_{\mathbf{B}}(y_k)\|_{\mathbf{B}}^4 \\ &= \frac{L_3}{12}\|x_{k+1} - T_{\mathbf{B}}(y_k)\|_{\mathbf{B}}^4,\end{aligned}$$

and so it follows from Corollary 18 that

$$\|x_{k+1} - T_{\mathbf{B}}(y_k)\|_{\mathbf{B}} \leq \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}.$$

■

F.5. Proof of Lemma 21.

Proof Let $\beta \stackrel{\text{def}}{=} x_{k+1} - T_{\mathbf{B}}(y_k)$. We have that

$$\begin{aligned}\left|\hat{r}(x_{k+1},y_k)^2 - r(y_k)^2\right| &= \left|\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2 - \|x_{k+1} - y_k\|_{\mathbf{B}}^2\right| \\ &= \left|\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2 - \|T_{\mathbf{B}}(y_k) + \beta - y_k\|_{\mathbf{B}}^2\right| \\ &= \left|\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2 + 2\langle\beta, \mathbf{B}(T_{\mathbf{B}}(y_k) - y_k)\rangle + \|\beta\|_{\mathbf{B}}^2 - \|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2\right| \\ &\leq 2\|\beta\|_{\mathbf{B}}\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}} + \|\beta\|_{\mathbf{B}}^2.\end{aligned}$$

Now, by Lemma 19, we know that $\|\beta\|_{\mathbf{B}} \leq \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}$, and so it follows from the definition of $\mathcal{P}$ that

$$\left|\hat{r}(x_{k+1},y_k)^2 - r(y_k)^2\right| \leq 6\left(\frac{\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/4}\mathcal{P}^{1/2} + \left(\frac{12\tilde{\varepsilon}_{aam}}{L_3}\right)^{1/2}.$$

■

F.6. Proof of Lemma 22.

Proof The lemma follows directly from Lemma 20, since

$$\begin{aligned}\langle\nabla f(x_{k+1}), y_k - x_{k+1}\rangle &\geq \frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) \\ &\quad - \frac{3Z(x_{k+1},y_k)W(x_{k+1},y_k)\tilde{\varepsilon}_{aam}^{1/4}}{L_3^{5/4}\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)} \\ &\geq \frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) - \frac{\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{c\rho_{\text{init}}^-},\end{aligned}$$

where we let

$$\mathcal{W} \stackrel{\text{def}}{=} \max_{x,y\in\mathscr{L}} Z(x,y)W(x,y). \tag{65}$$

■

F.7. Proof of Lemma 23.

Proof By sufficiently small, we mean that $\tilde{\varepsilon}_{aam} > 0$ is chosen such that

$$\tilde{\varepsilon}_{aam} \leq \min\left\{\left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{Q}}\right)^4, \left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{W}}\right)^4, \frac{1}{2}\right\},$$

where $\tilde{\varepsilon}_{rs}$ is as defined in the algorithm.

Following the standard line of reasoning, as presented by Nesterov (2018b), we proceed via proof by induction. For $k = 0$,

$$A_0 f(x_0) + B_0 = \min_{x\in\mathbb{R}^d} \psi_0 = 0, \quad f(x_0) \leq \mathcal{F}, \quad \|v_0 - x^*\|_{\mathbf{B}}^2 = \|x_0 - x^*\|_{\mathbf{B}}^2, \quad \text{and} \quad v_0 = x_0 \in \mathscr{L}.$$

Now suppose, for some $k \geq 0$, that (35) and (36) hold. To show that $\rho_{\text{init}}^+ = \mathcal{P}$ is a valid upper bound on ρ_k^*, we note that for any $\tau \in [0,1]$, letting $y_k = (1-\tau)x_k + \tau v_k$, we have that $f(y_k) \leq \max\{f(x_k), f(v_k)\} \leq \max\{\mathcal{F}, f(v_k)\}$, by our inductive assumption. We also know by our inductive assumption that $\|v_k - x^*\|_{\mathbf{B}}^2 \leq \|x_0 - x^*\|_{\mathbf{B}}^2$. Thus, since

$$\|v_k - x_0\|_{\mathbf{B}}^2 \leq 2\|v_k - x^*\|_{\mathbf{B}}^2 + 2\|x_0 - x^*\|_{\mathbf{B}}^2 \leq 4\|x_0 - x^*\|_{\mathbf{B}}^2,$$

it follows that $v_k \in \mathcal{K}$, which means that $f(v_k) \leq \mathcal{F}$, and so $f(y_k) \leq \mathcal{F}$. We then have that, for all $\tau \in [0,1]$, $\|T_{\mathbf{B}}(y_k) - y_k\|_{\mathbf{B}}^2 \leq \mathcal{P}$, where $\mathcal{P}$ is defined as in (21), since $f(T_{\mathbf{B}}(y_k)) \leq f(y_k)$ for all $x \in \mathbb{R}^d$. Thus, $\mathcal{P}$ is a valid upper bound on ρ_k^*.

For the lower bound on ρ_k^*, we note that, based on the condition for when the RhoSearch procedure is reached in FastQuartic, it must be the case that $\rho_{\text{init}}^- \leq (1+\tilde{\varepsilon}_{fs})\|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2$ and $\rho_{\text{init}}^- \leq \|x_{k+1}^- - y_k^-\|_{\mathbf{B}}^2 - \mathcal{Q}\tilde{\varepsilon}_{aam}^{1/4}$. Thus, from (47), it can be seen that $\rho_{\text{init}}^- \leq \zeta(\rho_{\text{init}}^-)$, and so it follows that $\rho_{\text{init}}^- \leq \rho_k^*$. Therefore, the correctness of RhoSearch can be ensured.

With this observation in hand, we may see that, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
\psi_{k+1}(x) &\geq \psi_k^* + \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 + a_{k+1}\left[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle\right] \\
&\geq A_k f(x_k) + B_k + \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 + a_{k+1}\left[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle\right] \\
&\geq A_k(f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle) + B_k + \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 \\
&\quad + a_{k+1}\left[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle\right] \\
&= A_{k+1} f(x_{k+1}) + B_k + \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 + \langle \nabla f(x_{k+1}), A_k(x_k - x_{k+1}) + a_{k+1}(x - x_{k+1})\rangle \\
&= A_{k+1} f(x_{k+1}) + B_k + \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 + \langle \nabla f(x_{k+1}), a_{k+1}(x - v_k) + A_{k+1}(y_k - x_{k+1})\rangle,
\end{aligned}$$

where the last equalities is due to the fact that $A_{k+1}y_k = A_k x_k + a_{k+1}v_k$. Note that

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|x - v_k\|_{\mathbf{B}}^2 + \langle \nabla f(x_{k+1}), a_{k+1}(x - v_k)\rangle = -\frac{a_{k+1}^2}{2}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2.$$

Combining this observation with Lemma 22, the fact that $\rho_{\text{init}}^- \leq \|x_{j+1}^- - y_j^-\|_{\mathbf{B}}^2$, and our choice of $\tilde{\varepsilon}_{aam}$, we have

$$\begin{aligned}
\min_{x\in\mathbb{R}^d}\psi_{k+1}(x) &\geq A_{k+1}f(x_{k+1}) + B_k - \frac{a_{k+1}^2}{2}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \langle \nabla f(x_{k+1}), A_{k+1}(y_k - x_{k+1})\rangle \\
&\geq A_{k+1}f(x_{k+1}) + B_k - \frac{A_{k+1}}{2L_3\rho_k}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 \\
&\quad + A_{k+1}\left(\frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) - \frac{\mathcal{W}\tilde{\varepsilon}_{aam}^{1/4}}{\rho_{\text{init}}^-}\right) \\
&\geq A_{k+1}f(x_{k+1}) + B_k - \frac{A_{k+1}}{2L_3\rho_k}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 \\
&\quad + A_{k+1}\left(\frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) - \frac{\tilde{\varepsilon}_{rs}^2}{1000\rho_{\text{init}}^-}\right).
\end{aligned}$$

We also know, by the guarantees of RhoSearch in Theorem 34, along with the choice of $\tilde{\varepsilon}_{aam}$, that $\rho_k \geq (1-\tilde{\varepsilon}_{rs})\hat{\zeta}(\rho_k) = (1-\tilde{\varepsilon}_{rs})\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)$, and so

$$\begin{aligned}
\min_{x\in\mathbb{R}^d}\psi_{k+1}(x) &\geq A_{k+1}f(x_{k+1}) + B_k - \frac{A_{k+1}}{2L_3(1-\tilde{\varepsilon}_{rs})\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 \\
&\quad + A_{k+1}\left(\frac{1}{2L_3\hat{r}_{\mathbf{B}}^2(x_{k+1},y_k)}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) - \frac{\tilde{\varepsilon}_{rs}^2}{1000\rho_{\text{init}}^-}\right) \\
&\geq A_{k+1}f(x_{k+1}) + B_k + A_{k+1}\left(\frac{3L_3}{8}\hat{r}_{\mathbf{B}}^4(x_{k+1},y_k) - \tilde{\varepsilon}_{curr}\right),
\end{aligned}$$

where

$$\tilde{\varepsilon}_{curr} \stackrel{\text{def}}{=} \frac{\tilde{\varepsilon}_{rs}}{2L_3(1-\tilde{\varepsilon}_{rs})\rho_{\text{init}}^-}\|\nabla f(x_{k+1})\|_{\mathbf{B}^{-1}}^2 + \frac{\tilde{\varepsilon}_{rs}^2}{1000\rho_{\text{init}}^-}.$$

Therefore, by our choice of $\tilde{\varepsilon}_{rs} \leq \frac{L_3(\rho_{\text{init}}^-)^3}{1000\mathcal{T}}$, where

$$\mathcal{T} \stackrel{\text{def}}{=} \frac{\mathcal{G}}{L_3} + \frac{1}{1000}, \tag{66}$$

(35) holds for $k+1$, proving the induction step. In addition, we may note that

$$\begin{aligned}
\psi_{k+1}(x) &= \frac{1}{2}\|x-x_0\|_{\mathbf{B}}^2 + \sum_{i=0}^{k+1} a_i\left[f(x_{k+1}) + \langle\nabla f(x_{k+1}), x - x_{k+1}\rangle\right] \\
&\leq \frac{1}{2}\|x-x_0\|_{\mathbf{B}}^2 + \sum_{i=0}^{k+1} a_i f(x) \\
&= A_{k+1}f(x) + \frac{1}{2}\|x-x_0\|_{\mathbf{B}}^2.
\end{aligned}$$

Since $v_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^d} \psi_{k+1}(x)$ and $\psi_{k+1}(x)$ is a quadratic function, it follows that, for all $x \in \mathbb{R}^d$,

$$\begin{aligned}\psi_{k+1}(x) &= \psi_{k+1}(v_{k+1}) + \langle \nabla\psi_{k+1}(v_{k+1}), x - v_{k+1}\rangle + \frac{1}{2}\|x - v_{k+1}\|_{\mathbf{B}}^2 \\ &= \psi_{k+1}(v_{k+1}) + \frac{1}{2}\|x - v_{k+1}\|_{\mathbf{B}}^2 \\ &\leq A_{k+1}f(x) + \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2.\end{aligned}$$

Taken together, this gives us that

$$\begin{aligned}A_{k+1}f(x_{k+1}) + B_{k+1} + \frac{1}{2}\|x - v_{k+1}\|_{\mathbf{B}}^2 &\leq \min_{x\in\mathbb{R}^d} \psi_{k+1}(x) + \frac{1}{2}\|x - v_{k+1}\|_{\mathbf{B}}^2 \\ &= \psi_{k+1}(v_{k+1}) + \frac{1}{2}\|x - v_{k+1}\|_{\mathbf{B}}^2 \\ &\leq A_{k+1}f(x) + \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2.\end{aligned}$$

Rearranging and letting $x = x^*$, we have that

$$A_{k+1}(f(x_{k+1}) - f(x^*)) + B_{k+1} + \frac{1}{2}\|x^* - v_{k+1}\|_{\mathbf{B}}^2 \leq \frac{1}{2}\|x^* - x_0\|_{\mathbf{B}}^2,$$

and so it follows that

$$\|v_{k+1} - x^*\|_{\mathbf{B}}^2 \leq \|x_0 - x^*\|_{\mathbf{B}}^2$$

and $v_{k+1}, x_{k+1} \in \mathscr{L}$. ■

F.8. Proof of Corollary 24.

Proof Note that, for all $k \geq 0$, $x \in \mathbb{R}^d$,

$$\begin{aligned}\psi_k(x) &= \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2 + \sum_{i=0}^{k} a_i \left[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle\right] \\ &\leq \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2 + \sum_{i=0}^{k} a_i f(x) \\ &= A_k f(x) + \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2,\end{aligned}$$

and so it follows from Lemma 23 that

$$A_k f(x_k) + B_k \leq \min_{x\in\mathbb{R}^d} \psi_k(x) \leq \min_{x\in\mathbb{R}^d} A_k f(x) + \frac{1}{2}\|x - x_0\|_{\mathbf{B}}^2 = A_k f(x^*) + \frac{1}{2}\|x^* - x_0\|_{\mathbf{B}}^2.$$

Rearranging, we have

$$\frac{3L_3}{16}\sum_{i=0}^{k-1} A_{i+1}\hat{r}_{\mathbf{B}}^4(x_{i+1}, y_i) = B_k \leq A_k(f(x^*) - f(x_k)) + \frac{1}{2}\|x^* - x_0\|_{\mathbf{B}}^2 \leq \frac{1}{2}\|x^* - x_0\|_{\mathbf{B}}^2$$

and so

$$f(x_k) - f(x^*) \leq \frac{1}{2A_k}\|x^* - x_0\|_{\mathbf{B}}^2.$$

■

F.9. Proof of Theorem 25

Proof By combining Corollary 24 with Lemma 36, we observe that

$$f(x_k) - f(x^*) \leq \frac{1}{2A_k}\|x_0 - x^*\|_{\mathbf{B}}^2 \leq \frac{128L_3\|x_0 - x^*\|_{\mathbf{B}}^4}{3}\left(\frac{2}{k+1}\right)^5.$$

■

F.10. Proof of Theorem 26.

Proof Let $\mathcal{Z} \overset{\text{def}}{=} \max\{\mathcal{A}, \mathcal{G}, \mathcal{P}, \mathcal{Q}, \mathcal{R}, \mathcal{V}, \mathcal{W}, L_3\}$. By appropriate initialization, we mean that ρ_{init}^-, $\tilde{\varepsilon}_{aam}$ are chosen such that $\rho_{\text{init}}^- \leq \frac{\varepsilon}{2L_3\mathcal{P}}$, and

$$\begin{aligned}\tilde{\varepsilon}_{aam} &< \min\left\{\left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{Q}}\right)^4, \left(\frac{\tilde{\varepsilon}_{rs}^2}{1000\mathcal{W}}\right)^4, \left(\frac{\tilde{\varepsilon}_{fs}}{\mathcal{V}(1+\tilde{\varepsilon}_{fs})}\right)^4, \left(\frac{\tilde{\varepsilon}_{fs}\rho_{\text{init}}^-}{\mathcal{Q}(1+\tilde{\varepsilon}_{fs})}\right)^4, \left(\frac{L_3(\rho_{\text{init}}^-)^3}{1000\mathcal{W}}\right)^4, \frac{1}{2}\right\} \\ &\leq \min\left\{O\left(poly\left(\frac{\varepsilon}{\mathcal{Z}}\right)\right), \frac{1}{2}\right\},\end{aligned}$$

where $\tilde{\varepsilon}_{fs}$ and $\tilde{\varepsilon}_{rs}$ are as defined in the FastQuartic algorithm. Thus, based on our choices of ρ_{init}^- and $\tilde{\varepsilon}_{aam}$, the iteration complexity follows immediately from Theorems 25 and 37. Each iteration of FastQuartic requires at most $O(\log(\frac{\mathcal{Z}}{\varepsilon}))$ iterations of RhoSearch, each of which requires at most $O(\log(\frac{\mathcal{Z}}{\varepsilon}))$ iterations of ApproxAuxMin, and each iteration of ApproxAuxMin requires at most $O(\log^{O(1)}(\frac{\mathcal{Z}}{\varepsilon}))$ calls to a gradient oracle and linear system solver. Taken together, this gives us a total computational cost of $O(\log^{O(1)}(\frac{\mathcal{Z}}{\varepsilon}))$ calls to a gradient oracle and linear system solver per iteration of FastQuartic. ■

F.11. Proofs of Theorems 12 and 13

To prove Theorems 12 and 13, we first rely on at Corollaries 27 and 28, found in Appendix B. Their proofs simply follow from Theorem 26, using the smoothness guarantees provided by Theorems 7 and 8, respectively. Thus, the proof of Theorem 12 follows by combining Fact 1, for $\mu = \frac{\varepsilon}{2\log(m)}$, with Corollary 27, while Theorem 13 follows from combining Lemma 4, for $\mu = \frac{\varepsilon}{4\lambda d}$, with Corollary 28.