A Proof of Main Results

In this section, we give formal proofs of Theorems 1 and 3. We start by providing proofs of several useful auxiliary lemmas.

Remark 3. It suffices to assume that $\epsilon \leq \frac{\ell^2}{\rho}$ for the following analysis, since otherwise every point $\mathbf{x}$ satisfies the second-order condition $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\epsilon}$ trivially by the Lipschitz-gradient assumption.

A.1 Set-Up and Notation

Here we remind the reader of the relevant notation and provide further background from Nesterov and Polyak [2006] on the cubic-regularized Newton method. We denote the stochastic gradient as

$$\mathbf{g}_t = \frac{1}{|S_1|} \sum_{\xi_i \in S_1} \nabla f(\mathbf{x}_t, \xi_i)$$

and the stochastic Hessian as

$$\mathbf{B}_t = \frac{1}{|S_2|} \sum_{\xi_i \in S_2} \nabla^2 f(\mathbf{x}_t, \xi_i),$$

both for iteration t. We draw a sufficient number of samples $|S_1|$ and $|S_2|$ so that the concentration conditions

$$\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\| \leq c_1 \cdot \epsilon,$$

$$\forall \mathbf{v}, \|(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t))\mathbf{v}\| \leq c_2 \cdot \sqrt{\rho\epsilon}\|\mathbf{v}\|.$$

are satisfied for sufficiently small c_1, c_2 (see Lemma 4 for more details). The cubic-regularized Newton subproblem is to minimize

$$m_t(\mathbf{y}) = f(\mathbf{x}_t) + (\mathbf{y} - \mathbf{x}_t)^\top \mathbf{g}_t + \frac{1}{2}(\mathbf{y} - \mathbf{x}_t)^\top \mathbf{B}_t(\mathbf{y} - \mathbf{x}_t) + \frac{\rho}{6}\|\mathbf{y} - \mathbf{x}_t\|^3. \tag{14}$$

We denote the global optimizer of $m_t(\cdot)$ as $\mathbf{x}_t + \mathbf{\Delta}_t^\star$, that is $\mathbf{\Delta}_t^\star = \operatorname{argmin}_z m_k(\mathbf{z} + \mathbf{x}_k)$.

As shown in Nesterov and Polyak [2006] a global optima of Equation (14) satisfies:

$$\mathbf{g}_t + \mathbf{B}_t \mathbf{\Delta}_t^\star + \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|\mathbf{\Delta}_t^\star = 0. \tag{15}$$

$$\mathbf{B}_t + \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\| I \succeq 0. \tag{16}$$

Equation (15) is the first-order stationary condition, while Equation (16) follows from a duality argument. In practice, we will not be able to directly compute $\mathbf{\Delta}_t^\star$ so will instead use a Cubic-Subsolver routine which must satisfy:

Condition 1. For any fixed, small constant c_3, c_4, Cubic-Subsolver$(\mathbf{g}, \mathbf{B}[\cdot], \epsilon)$ terminates within $\mathcal{T}(\epsilon)$ gradient iterations (which may depend on c_3, c_4), and returns a $\mathbf{\Delta}$ satisfying at least one of the following:

1. $\max\{\tilde{m}(\mathbf{\Delta}), f(\mathbf{x}_t + \mathbf{\Delta}) - f(\mathbf{x}_t)\} \leq -\Omega(\sqrt{\epsilon^3/\rho})$. **(Case 1)**
2. $\|\mathbf{\Delta}\| \leq \|\mathbf{\Delta}^\star\| + c_4\sqrt{\frac{\epsilon}{\rho}}$ and, if $\|\mathbf{\Delta}^\star\| \geq \frac{1}{2}\sqrt{\epsilon/\rho}$, then $\tilde{m}(\mathbf{\Delta}) \leq \tilde{m}(\mathbf{\Delta}^\star) + \frac{c_3}{12} \cdot \rho\|\mathbf{\Delta}^\star\|^3$. **(Case 2)**

A.2 Auxiliary Lemmas

We begin by providing the proof of several useful auxiliary lemmas. First we provide the proof of Lemma 4 which characterize the concentration conditions.

Lemma 4. *For any fixed small constants c_1, c_2, we can pick gradient and Hessian mini-batch sizes $n_1 = \tilde{\mathcal{O}}\left(\max\left(\frac{M_1}{\epsilon}, \frac{\sigma_1^2}{\epsilon^2}\right)\right)$ and $n_2 = \tilde{\mathcal{O}}\left(\max\left(\frac{M_2}{\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{\rho\epsilon}\right)\right)$ so that with probability $1-\delta'$,*

$$\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\| \le c_1 \cdot \epsilon, \tag{12}$$

$$\forall \mathbf{v}, \|(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t))\mathbf{v}\| \le c_2 \cdot \sqrt{\rho\epsilon}\|\mathbf{v}\|. \tag{13}$$

Proof. We can use the matrix Bernstein inequality from Tropp et al. [2015] to control both the fluctuations in the stochastic gradients and stochastic Hessians under Assumption 2.

Recall that the spectral norm of a vector is equivalent to its vector norm. So the matrix variance of the centered gradients $\tilde{\mathbf{g}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\tilde{\nabla} f(\mathbf{x}, \xi_i)\right) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\nabla f(\mathbf{x}, \xi_i) - \nabla f(\mathbf{x})\right)$ is:

$$v[\tilde{\mathbf{g}}] = \frac{1}{n_1^2}\max\left\{\left\|\mathbb{E}\left[\sum_{i=1}^{n_1}\tilde{\nabla} f(\mathbf{x}, \xi_i)\tilde{\nabla} f(\mathbf{x}, \xi_i)^\top\right]\right\|, \left\|\mathbb{E}\left[\sum_{i=1}^{n_1}\tilde{\nabla} f(\mathbf{x}, \xi_i)^\top\tilde{\nabla} f(\mathbf{x}, \xi_i)\right]\right\|\right\} \le \frac{\sigma_1^2}{n_1}$$

using the triangle inequality and Jensens inequality. A direct application of the matrix Bernstein inequality gives:

$$\mathbb{P}\left[\|\mathbf{g} - \nabla f(\mathbf{x})\| \ge t\right] \le 2d\exp\left(-\frac{t^2/2}{v[\tilde{\mathbf{g}}] + M_1/(3n_1)}\right) \le 2d\exp\left(-\frac{3n_1}{8}\min\left\{\frac{t}{M_1}, \frac{t^2}{\sigma_1^2}\right\}\right) \implies$$
$$\|\mathbf{g} - \nabla f(\mathbf{x})\| \le t \text{ with probability } 1-\delta' \text{ for } n_1 \ge \max\left(\frac{M_1}{t}, \frac{\sigma_1^2}{t^2}\right)\frac{8}{3}\log\frac{2d}{\delta'}$$

Taking $t = c_1\epsilon$ gives the result.

The matrix variance of the centered Hessians $\tilde{\mathbf{B}} = \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\tilde{\nabla}^2 f(\mathbf{x}, \xi_i)\right) = \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\nabla^2 f(\mathbf{x}, \xi_i) - \nabla^2 f(\mathbf{x})\right)$, which are symmetric, is:

$$v[\tilde{\mathbf{B}}] = \frac{1}{n_2^2}\left\|\sum_{i=1}^{n_2}\mathbb{E}\left[\left(\tilde{\nabla}^2 f(\mathbf{x}, \xi_i)\right)^2\right]\right\| \le \frac{\sigma_2^2}{n_2} \tag{17}$$

once again using the triangle inequality and Jensens inequality. Another application of the matrix Bernstein inequality gives that:

$$\mathbb{P}[\|\mathbf{B} - \nabla^2 f(\mathbf{x}))\| \ge t] \le 2d\exp\left(-\frac{3n_2}{8}\min\{\frac{t}{M_2}, \frac{t^2}{\sigma_2^2}\}\right) \implies$$
$$\|\mathbf{B} - \nabla^2 f(\mathbf{x}))\| \le t \text{ with probability } 1-\delta' \text{ for } n_2 \ge \max(\frac{M_2}{t}, \frac{\sigma_2^2}{t^2})\frac{8}{3}\log\frac{2d}{\delta'}$$

Taking $t = c_2\sqrt{\rho\epsilon}$ ensures that the stochastic Hessian-vector products are controlled uniformly over $\mathbf{v}$:

$$\|(\mathbf{B} - \nabla^2 f(\mathbf{x}))\mathbf{v}\| \le c_2 \cdot \sqrt{\rho\epsilon}\|\mathbf{v}\|$$

using n_2 samples with probability $1-\delta'$.

□

Next we show Lemma 5 which will relate the change in the cubic function value to the norm $\|\boldsymbol{\Delta}_t^\star\|$.

Lemma 5. *Let m_t and $\boldsymbol{\Delta}_t^\star$ be defined as above. Then for all t,*

$$m_t(\mathbf{x}_t + \boldsymbol{\Delta}_t^\star) - m_t(\mathbf{x}_t) \le -\frac{1}{12}\rho\|\boldsymbol{\Delta}_t^\star\|^3.$$

Proof. Using the global optimality conditions for Equation (14) from Nesterov and Polyak [2006], we have the global optima $\mathbf{x}_t + \boldsymbol{\Delta}_t^\star$, satisfies:

$$\mathbf{g}_t + \mathbf{B}_t(\boldsymbol{\Delta}_t^\star) + \frac{\rho}{2}\|\boldsymbol{\Delta}_t^\star\|(\boldsymbol{\Delta}_t^\star) = 0$$

$$\mathbf{B}_t + \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\| I \succeq 0.$$

Together these conditions also imply that:

$$(\mathbf{\Delta}_t^\star)^\top \mathbf{g}_t + (\mathbf{\Delta}_t^\star)^\top \mathbf{B}_t(\mathbf{\Delta}_t^\star) + \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|^3 = 0$$
$$(\mathbf{\Delta}_t^\star)^\top \mathbf{B}_t(\mathbf{\Delta}_t^\star) + \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|^3 \geq 0.$$

Now immediately from the definition of stochastic cubic submodel model and the aforementioned conditions we have that:

$$\begin{aligned}
f(\mathbf{x}_t) - m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star) &= -(\mathbf{\Delta}_t^\star)^\top \mathbf{g}_t - \frac{1}{2}(\mathbf{\Delta}_t^\star)^\top \mathbf{B}_t(\mathbf{\Delta}_t^\star) - \frac{\rho}{6}\|\mathbf{x}_t + \mathbf{\Delta}_t^\star\|^3 \\
&= \frac{1}{2}(\mathbf{\Delta}_t^\star)^\top \mathbf{B}_t(\mathbf{\Delta}_t^\star) + \frac{1}{3}\rho\|\mathbf{\Delta}_t^\star\|^3 \\
&\geq \frac{1}{12}\rho\|\mathbf{\Delta}_t^\star\|^3
\end{aligned}$$

An identical statement appears as Lemma 10 in Nesterov and Polyak [2006], so this is merely restated here for completeness. □

Thus to guarantee sufficient descent it suffices to lower bound the $\|\mathbf{\Delta}_t^\star\|$. We now prove Lemma 6, which guarantees the sufficient "movement" for the exact update: $\|\mathbf{\Delta}_t^\star\|$. In particular this will allow us to show that when $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is not an ϵ-second-order stationary point then $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$.

Lemma 6. *Under the setting of Lemma 4 with sufficiently small constants c_1, c_2,*

$$\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\max\left\{\sqrt{\frac{1}{\rho}\left(\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| - \frac{\epsilon}{4}\right)}, \frac{1}{\rho}\left(\lambda_{\min}(\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)) - \frac{\sqrt{\rho\epsilon}}{4}\right)\right\}.$$

Proof. As a consequence of the global optimality condition, given in Equation (15), we have that:

$$\|\mathbf{g}_t + \mathbf{B}_t(\mathbf{\Delta}_t^\star)\| = \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|^2. \tag{18}$$

Moreover, from the Hessian-Lipschitz condition it follows that:

$$\left\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{\Delta}_t^\star)\right\| \leq \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|^2. \tag{19}$$

Combining the concentration assumptions with Equation (18) and Inequality (19), we obtain:

$$\begin{aligned}
\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| &= \left\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{\Delta}_t^\star)\right\| + \left\|\nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{\Delta}_t^\star)\right\| \\
&\leq \left\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)(\mathbf{\Delta}_t^\star)\right\| + \|\mathbf{g}_t + \mathbf{B}_t(\mathbf{\Delta}_t^\star)\| \\
&+ \|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\| + \left\|(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t))\mathbf{\Delta}_t^\star\right\| \\
&\leq \rho\|\mathbf{\Delta}_t^\star\|^2 + c_1\epsilon + c_2\sqrt{\rho\epsilon}\|\mathbf{\Delta}_t^\star\|.
\end{aligned} \tag{20}$$

An application of the Fenchel-Young inequality to the final term in Equation (20) then yields:

$$\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| \leq \rho(1 + \frac{c_2}{2})\|\mathbf{\Delta}_t^\star\|^2 + (c_1 + \frac{c_2}{2})\epsilon \implies$$
$$\frac{1}{\rho(1 + \frac{c_2}{2})}\left(\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| - (c_1 + \frac{c_2}{2})\epsilon\right) \leq \|\mathbf{\Delta}_t^\star\|^2,$$

which lower bounds $\|\mathbf{\Delta}_t^\star\|$ with respect to the gradient at $\mathbf{x}_t$. For the corresponding Hessian lower bound we first utilize the Hessian Lipschitz condition:

$$\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) \succeq \nabla^2 f(\mathbf{x}_t) - \rho\|\mathbf{\Delta}_t^\star\| I$$

$$\succeq \mathbf{B}_t - c_2\sqrt{\rho\epsilon}I - \rho\|\mathbf{\Delta}_t^\star\|I$$
$$\succeq -c_2\sqrt{\rho\epsilon}I - \frac{3}{2}\rho\|\mathbf{\Delta}_t^\star\|I,$$

followed by the concentration condition and the optimality condition (16). This immediately implies

$$\|\mathbf{\Delta}_t^\star\|I \succeq -\frac{2}{3\rho}\left(\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) + c_2\sqrt{\rho\epsilon}I\right) \implies$$
$$\|\mathbf{\Delta}_t^\star\| \geq -\frac{2}{3\rho}\lambda_{\min}(\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)) - \frac{2c_2}{3\sqrt{\rho}}\sqrt{\epsilon}$$

Combining we obtain that:

$$\|\mathbf{\Delta}_t^\star\| \geq \max\left\{\sqrt{\frac{1}{\rho(1+\frac{c_2}{2})}\left(\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| - (c_1 + \frac{c_2}{2})\epsilon\right)}, -\frac{2}{3\rho}\lambda_n(\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)) - \frac{2c_2}{3\sqrt{\rho}}\sqrt{\epsilon}\right\}.$$

We consider the case of large gradient and large Hessian in turn (one of which must hold since $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is not an ϵ-second-order stationary point). There exist c_1, c_2 in the following so that we can obtain:

- If $\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| > \epsilon$, then we have that

$$\|\mathbf{\Delta}_t^\star\| > \sqrt{\frac{1}{\rho(1+\frac{c_2}{2})}\left(\|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| - (c_1 + \frac{c_2}{2})\epsilon\right)} \geq \sqrt{\frac{1-c_1-\frac{c_2}{2}}{1+\frac{c_2}{2}}}\sqrt{\frac{\epsilon}{\rho}} > \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}. \quad (21)$$

- If $-\lambda_n(\nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)) > \sqrt{\rho\epsilon}$, then we have that $\|\mathbf{\Delta}_t^\star\| > \frac{2}{3}\sqrt{\frac{\epsilon}{\rho}} - \frac{2c_2}{3}\sqrt{\frac{\epsilon}{\rho}} = \frac{2}{3}(1-c_2)\sqrt{\frac{\epsilon}{\rho}} > \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$.

We can similarly check the lower bounds directly stated are true. Choosing $c_1 = \frac{1}{200}$ and $c_2 = \frac{1}{200}$ will verify these inequalities for example. □

A.3 Proof of Claim 1

Here we provide a proof of statement equivalent to Claim 1 in the full, non-stochastic setting with approximate model minimization. We focus on the case when the Cubic-Subsolver routine executes **Case 2**, since the result is vacuously true when the routine executes **Case 1**. Our first lemma will both help show sufficient descent and provide a stopping condition for Algorithm 1. For context, recall that when $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is not an ϵ-second-order stationary point then $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$ by Lemma 6.

Lemma 7. *If the routine Cubic-Subsolver uses **Case 2**, and if $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$, then it will return a point $\mathbf{\Delta}$ satisfying $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) \leq m_t(\mathbf{x}_t) - \frac{1-c_3}{12}\rho\|\mathbf{\Delta}_t^\star\|^3 \leq \frac{1-c_3}{96}\sqrt{\frac{\epsilon^3}{\rho}}$.*

Proof. In the case when $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$, by the definition of the routine Cubic-Subsolver we can ensure that $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) \leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star) + \frac{c_3}{12}\rho\|\mathbf{\Delta}_t^\star\|^3$ for arbitarily small c_3 using $\mathcal{T}(\epsilon)$ iterations. We can now combine the aforementioned display with Lemma 5 (recalling that $m_t(\mathbf{x}_t) = f(\mathbf{x}_t)$) to conclude that:

$$m_t(\mathbf{x}_t + \mathbf{\Delta}_t) \leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star) + \frac{c_3}{12}\rho\|\mathbf{\Delta}_t^\star\|^3$$
$$m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star) \leq m_t(\mathbf{x}_t) - \frac{\rho}{12}\|\mathbf{\Delta}_t^\star\|^3 \implies \quad (22)$$
$$m_t(\mathbf{x}_t + \mathbf{\Delta}_t) \leq m_t(\mathbf{x}_t) - (\frac{1-c_3}{12})\rho\|\mathbf{\Delta}_t^\star\|^3 \leq m_t(\mathbf{x}_t) - \frac{(1-c_3)}{96}\sqrt{\frac{\epsilon^3}{\rho}}. \quad (23)$$

for suitable choice of c_3 which can be made arbitarily small. □

Claim 1. *Assume we are in the setting of Lemma 4 with sufficiently small constants c_1, c_2. If $\mathbf{\Delta}$ is the output of the routine Cubic-Subsolver when executing **Case 2** and if $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is not an ϵ-second-order stationary point of f, then $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \le -\frac{1-c_3}{96}\sqrt{\frac{\epsilon^3}{\rho}}$.*

Proof. This is an immediate consequence of Lemmas 6 and 7. □

If we do not observe sufficient descent in the cubic submodel (which is not possible in **Case 1** by definition) then as a consequence of Claim 1 and Lemma 7 we can conclude that $\|\mathbf{\Delta}_t^\star\| \le \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$ and that $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ *is* an ϵ-second-order stationary point. However, we cannot compute $\mathbf{\Delta}_t^\star$ directly. So instead we use a final gradient descent loop in Algorithm 2, to ensure the final point returned in this scenario will be an ϵ-second-order stationary point up to a rescaling.

Lemma 8. *Assume we are in the setting of Lemma 4 with sufficiently small constants c_1, c_2. If $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is an ϵ-second-order stationary point of f, and $\|\mathbf{\Delta}_t^\star\| \le \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$, then Algorithm 2 will output a point $\mathbf{\Delta}$ such that $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{\Delta}$ is a 4ϵ-second-order stationary point of f.*

Proof. Since $\mathbf{x}_t + \mathbf{\Delta}_t^\star$ is an ϵ-second order stationary point of f, by gradient smoothness and the concentration conditions we have that $\|\mathbf{g}_t\| \le \|\nabla f(\mathbf{x}_t + \mathbf{\Delta}_t^\star)\| + \ell\|\mathbf{\Delta}_t^\star\| + \|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\| \le (1+c_1)\epsilon + \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}\ell \le (\frac{3}{2} + 1 + c_1)\frac{\ell^2}{\rho} \le \frac{19}{16}\frac{\ell^2}{\rho}$ for sufficiently small c_1. Then we can verify the step-size choice $\eta = \frac{1}{20}\ell$ and initialization at $\mathbf{\Delta} = 0$ (in the centered coordinates) for the routine Cubic-FinalSubsolver verifies Assumptions A and B[2] in Carmon and Duchi [2016]. So, by Corollary 2.5 in Carmon and Duchi [2016]—which states the norms of the gradient descent iterates, $\|\mathbf{\Delta}\|$, are non-decreasing and satisfy $\|\mathbf{\Delta}\| \le \|\mathbf{\Delta}_t^\star\|$—we have that $\|\mathbf{\Delta} - \mathbf{\Delta}_t^\star\| \le 2\|\mathbf{\Delta}_t^\star\| \le \sqrt{\frac{\epsilon}{\rho}}$.

We first show that $-\lambda_{\min}(\nabla^2 f(\mathbf{x}_{t+1})) \lesssim \sqrt{\rho\epsilon}$. Since f is ρ-Hessian-Lipschitz we have that:

$$\nabla^2 f(\mathbf{x}_{t+1}) \succeq \nabla^2 f(\mathbf{x}_t + \mathbf{\Delta}_t^\star) - \rho 2\|\mathbf{\Delta}_t^\star\| I \succeq -2\sqrt{\rho\epsilon} I.$$

We now show that $\|\nabla f(\mathbf{x}_{t+1})\| \lesssim \epsilon$ and thus also small. Once again using that f is ρ-Hessian-Lipschitz (Lemma 1 in Nesterov and Polyak [2006]) we have that:

$$\left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)\mathbf{\Delta}\right\| \le \frac{\rho}{2}\|\mathbf{\Delta}\|^2 \le \frac{\rho}{2}\|\mathbf{\Delta}_t^\star\|^2 \le \frac{\epsilon}{8}.$$

Now, by the termination condition in Algorithm 2 we have that $\left\|\mathbf{g} + \mathbf{B}\mathbf{\Delta} + \frac{\rho}{2}\|\mathbf{\Delta}\|\mathbf{\Delta}\right\| < \frac{\epsilon}{2}$. So,

$$\|\mathbf{g} + \mathbf{B}\mathbf{\Delta}\| < \frac{\epsilon}{2} + \frac{\rho}{2}\|\mathbf{\Delta}\|^2 \le \frac{5}{8}\epsilon.$$

Using gradient/Hessian concentration with the previous displays we also obtain that:

$$\begin{aligned} &\|\nabla f(\mathbf{x}_{t+1})\| - \|\mathbf{g} - \nabla f(\mathbf{x}_t)\| - \left\|(\mathbf{B} - \nabla^2 f(\mathbf{x}_t))\mathbf{\Delta}\right\| - \|\mathbf{g} + \mathbf{B}\mathbf{\Delta}\| \le \left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)\mathbf{\Delta}\right\| \\ &\implies \|\nabla f(\mathbf{x}_{t+1})\| \le \left(c_1 + \frac{c_2}{2} + \frac{5}{8} + \frac{1}{8}\right)\epsilon \le \epsilon, \end{aligned}$$

for sufficiently small c_1 and c_2.

Let us now bound the iteration complexity of this step. From our previous argument we have that $\|\mathbf{g}_t\| \le (1+c_1)\epsilon + \frac{\ell}{2\sqrt{\rho}}\sqrt{\epsilon}$. Similarly, the concentration conditions imply $\|\mathbf{B}_t\mathbf{\Delta}_t^\star\| \le (\ell + c_2\sqrt{\rho\epsilon})\|\mathbf{\Delta}_t^\star\|$. Thus we have that $m_t(\mathbf{x}_t) - m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star) = ((1+c_1)\epsilon + \frac{\ell}{2\sqrt{\rho}}\sqrt{\epsilon})\|\mathbf{\Delta}_t^\star\| + \frac{1}{2}(\ell + c_2\sqrt{\rho\epsilon})\|\mathbf{\Delta}_t^\star\|^2 + \frac{\rho}{6}\|\mathbf{\Delta}_t^\star\|^3 \le \frac{3\ell}{\rho}\epsilon + \left(\frac{1+c_1+4c_2}{8} + \frac{1}{48}\right)\sqrt{\frac{\epsilon^3}{\rho}} \le \mathcal{O}(1)\cdot\frac{\epsilon\ell}{\rho}$ since c_1, c_2 are numerical constants that can be made arbitrarily small.

So by the standard analysis of gradient descent for smooth functions, see Nesterov [2013] for example, we have that Algorithm 2 will terminate in at most $\le \lceil \frac{m_t(\mathbf{x}_t) - m_t(\mathbf{x}_t + \mathbf{\Delta}_t^\star)}{\eta(\epsilon/2)^2} \rceil \le \mathcal{O}(1)\cdot(\frac{\ell^2}{\rho\epsilon})$ iterations. This will take at most $\tilde{\mathcal{O}}(\max(\frac{M_1}{\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{\epsilon})\cdot\frac{\ell^2}{\rho\epsilon})$ Hessian-vector products and $\tilde{\mathcal{O}}(\max(\frac{M_1}{\epsilon}, \frac{\sigma_1^2}{\epsilon^2}))$ gradient evaluations which will be subleading in the overall complexity. □

[2]See Appendix Section B.2 for more details.

A.4 Proof of Claim 2

We now prove our main descent lemma equivalent to **Claim 2**—this will show if the cubic submodel has a large decrease, then the underlying true function must also have large decrease. As before we focus on the case when the Cubic-Subsolver routine executes **Case 2** since the result is vacuously true in **Case 1**.

Claim 2. *Assume we are in the setting of Lemma 4 with sufficiently small constants c_1, c_2. If the Cubic-Subsolver routine uses* ***Case 2****, and if* $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \leq -(\frac{1-c_3}{96})\sqrt{\frac{\epsilon^3}{\rho}}$*, then* $f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq -\left(\frac{1-c_3-c_5}{96}\right)\sqrt{\frac{\epsilon^3}{\rho}}$.

Proof. Using that f is ρ-Hessian Lipschitz (and hence admits a cubic majorizer by Lemma 1 in Nesterov and Polyak [2006] for example) as well as the concentration conditions we have that:

$$
\begin{aligned}
&f(\mathbf{x}_t + \mathbf{\Delta}_t) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{\Delta}_t + \frac{1}{2}\mathbf{\Delta}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{\Delta}_t + \frac{\rho}{6}\|\mathbf{\Delta}_t\|_2^3 \implies \\
&f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + (\nabla f(\mathbf{x}_t) - \mathbf{g}_t)^\top \mathbf{\Delta}_t + \frac{1}{2}\mathbf{\Delta}_t^\top(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t))\mathbf{\Delta}_t \\
&\leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + c_1\epsilon\|\mathbf{\Delta}_t\| + \frac{c_2}{2}\sqrt{\rho\epsilon}\|\mathbf{\Delta}_t\|^2 \\
&\leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + c_1\epsilon\left(\|\mathbf{\Delta}_t^\star\| + c_4\sqrt{\frac{\epsilon}{\rho}}\right) + \frac{c_2}{2}\sqrt{\rho\epsilon}\left(\|\mathbf{\Delta}_t^\star\| + c_4\sqrt{\frac{\epsilon}{\rho}}\right)^2 \\
&\leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + (c_1 + c_2c_4)\epsilon\|\mathbf{\Delta}_t^\star\| + \frac{c_2c_4^2}{2}\sqrt{\rho\epsilon}\|\mathbf{\Delta}_t^\star\|^2 + (c_1 + \frac{c_2c_4}{2})c_4\sqrt{\frac{\epsilon^3}{\rho}},
\end{aligned} \tag{24}
$$

since by the definition the Cubic-Subsolver routine, when we use **Case 2** we have that $\|\mathbf{\Delta}_t\| \leq \|\mathbf{\Delta}_t^\star\| + c_4\sqrt{\frac{\epsilon}{\rho}}$. We now consider two different situations – when $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$ and when $\|\mathbf{\Delta}_t^\star\| \leq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$.

First, if $\|\mathbf{\Delta}_t^\star\| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$ then by Lemma 7 we may assume the stronger guarantee that $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \leq -(\frac{1-c_3}{12})\rho\|\mathbf{\Delta}_t^\star\|^3$. So by considering the above display in Equation (24) we can conclude that:

$$
\begin{aligned}
&f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + (c_1 + c_2c_4)\epsilon\|\mathbf{\Delta}_t^\star\| + \frac{c_2c_4^2}{2}\sqrt{\rho\epsilon}\|\mathbf{\Delta}_t^\star\|^2 + (c_1 + \frac{c_2c_4}{2})c_4\sqrt{\frac{\epsilon^3}{\rho}} \\
&\leq -\left(\frac{1 - c_3 - 48(c_1 + c_2c_4) - 12c_2c_4^2}{12}\right)\rho\|\mathbf{\Delta}_t^\star\|^3 + \left(c_1 + \frac{c_2c_4}{2}\right)c_4\sqrt{\frac{\epsilon^3}{\rho}} \\
&\leq -\left(\frac{1 - c_3 - 48c_1 - 48c_2c_4 - 96c_1c_4 - 60c_2c_4^2}{96}\right)\sqrt{\frac{\epsilon^3}{\rho}},
\end{aligned}
$$

since the numerical constants c_1, c_2, c_3 can be made arbitrarily small.

Now, if $\|\mathbf{\Delta}_t^\star\| \leq \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}}$, we directly use the assumption that $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \leq -(\frac{1-c_3}{96})\sqrt{\frac{\epsilon^3}{\rho}}$. Combining with the display in in Equation (24) we can conclude that:

$$
\begin{aligned}
&f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) + (c_1 + c_2c_4)\epsilon\|\mathbf{\Delta}_t^\star\| + \frac{c_2c_4^2}{2}\sqrt{\rho\epsilon}\|\mathbf{\Delta}_t^\star\|^2 + (c_1 + \frac{c_2c_4}{2})c_4\sqrt{\frac{\epsilon^3}{\rho}} \\
&\leq -\left(\frac{1-c_3}{96}\sqrt{\frac{\epsilon^3}{\rho}}\right) + \left((c_1 + c_2c_4)\epsilon \cdot \frac{1}{2}\sqrt{\frac{\epsilon}{\rho}} + \frac{c_2c_4^2}{2}\sqrt{\rho\epsilon} \cdot \frac{1}{4}\frac{\epsilon}{\rho} + (c_1 + \frac{c_2c_4}{2})c_4\sqrt{\frac{\epsilon^3}{\rho}}\right) \\
&\leq -\left(\frac{1 - c_3 - 48c_1 - 48c_2c_4 - 96c_1c_4 - 60c_2c_4^2}{96}\right)\sqrt{\frac{\epsilon^3}{\rho}},
\end{aligned}
$$

since the numerical constants c_1, c_2, c_3 can be made arbitrarily small. Indeed, recall that c_1 is the gradient concentration constant, c_2 is the Hessian-vector product concentration constant, and c_3 is the tolerance of the Cubic-Subsolver routine when using **Case 2**. Thus, in both situations, we have that:

$$f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq \frac{1 - c_3 - c_5}{96}\sqrt{\frac{\epsilon^3}{\rho}}, \tag{25}$$

denoting $c_5 = 48c_1 - 48c_2c_4 - 96c_1c_4 - 60c_2c_4^2$ for notational convenience (which can also be made arbitrarily small for sufficiently small c_1, c_2). $\square$

A.5 Proof of Theorem 1

We now prove the correctness of Algorithm 1. We assume, as usual, the underlying function $f(x)$ possesses a lower bound f^*.

Theorem 1. *There exists an absolute constant c such that if $f(\mathbf{x})$ satisfies Assumptions 1, 2, CubicSubsolver satisfies Condition 1 with c, $n_1 \geq \max(\frac{M_1}{c\epsilon}, \frac{\sigma_1^2}{c^2\epsilon^2})\log\left(\frac{d\sqrt{\rho}\Delta_f}{\epsilon^{1.5}\delta c}\right)$, and $n_2 \geq \max(\frac{M_2}{c\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{c^2\rho\epsilon})\log\left(\frac{d\sqrt{\rho}\Delta_f}{\epsilon^{1.5}\delta c}\right)$, then for all $\delta > 0$ and $\Delta_f \geq f(\mathbf{x}_0) - f^*$, Algorithm 1 will output an ϵ-second-order stationary point of f with probability at least $1 - \delta$ within*

$$\tilde{\mathcal{O}}\left(\frac{\sqrt{\rho}\Delta_f}{\epsilon^{1.5}}\left(\max\left(\frac{M_1}{\epsilon}, \frac{\sigma_1^2}{\epsilon^2}\right) + \max\left(\frac{M_2}{\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{\rho\epsilon}\right)\cdot\mathcal{T}(\epsilon)\right)\right) \tag{8}$$

total stochastic gradient and Hessian-vector product evaluations.

Proof. For notational convenience let **Case 1** of the routine Cubic Subsolver satisfy:

$$\max\{f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t), m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t)\} \leq -K_1\sqrt{\frac{\epsilon^3}{\rho}}.$$

and use $K_2 = \frac{1-c_3}{96}$ to denote the descent constant of the cubic submodel in the assumption of Claim 2. Further, let $K_{\text{prog}} = \min\{\frac{1-c_3-c_5}{96}, K_1\}$ which we will use as the progress constant corresponding to descent in the underlying function f. Without loss of generality, we assume that $-K_1 \leq -K_2$ for convenience in the proof. If $-K_1 \geq -K_2$, we can simply rescale the descent constant corresponding to **Case 2** for the cubic submodel, $\frac{1-c_3}{96}$, to be equal to $-K_1$, which will require shrinking c_1, c_2 proportionally to ensure that the rescaled version of the function descent constant, $\frac{1-c_3-c_5}{96}$, is positive.

Now, we choose c_1, c_2, c_3 so that $K_2 > 0$, $K_{\text{prog}} > 0$, and Lemma 6 holds in the aforementioned form. For the correctness of Algorithm 1 we choose the numerical constant in Line 7 as K_2 – so the "if statement" checks the condition $\Delta m = m_t(\mathbf{x}_{t+1}) - m_t(\mathbf{x}_t) \geq -K_2\sqrt{\frac{\epsilon'^3}{\rho}}$. Here we use a rescaled $\epsilon' = \frac{1}{4}\epsilon$ for the duration of the proof.

At each iteration the event that the setting of Lemma 4 hold has probability greater then $1 - 2\delta'$. Conditioned on this event let the routine Cubic-Subsolver have a further probability of at most δ' of failure. We now proceed with our analysis deterministically conditioned on the event E – that at each iteration the concentration conditions hold and the routine Cubic-Subsolver succeeds – which has probability greater then $1 - 3\delta' T_{\text{outer}} \geq 1 - \delta$ by a union bound for $\delta' = \frac{\delta}{3T_{\text{outer}}}$.

Let us now bound the iteration complexity of Algorithm 1 as T_{outer}. We cannot have the "if statement" in Line 7 fail indefinitely. At a given iteration, if the routine Cubic-Subsolver outputs a point $\mathbf{\Delta}$ that satisfies

$$m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \leq -K_2\sqrt{\frac{\epsilon'^3}{\rho}}$$

then by Claim 2 and the definition of **Case 1** of the Cubic-Subsolver we also have that:

$$f(\mathbf{x}_t + \mathbf{\Delta}_t) - f(\mathbf{x}_t) \leq -K_{\text{prog}}\sqrt{\frac{\epsilon'^3}{\rho}}.$$

Note if the Cubic-Subsolver uses **Case 1** in this iteration then we will vacuously achieve descent in both the underlying function f, and descent in the cubic submodel greater $-K_1\sqrt{\frac{\epsilon'^3}{\rho}}$. Since $-K_1\sqrt{\frac{\epsilon'^3}{\rho}} \leq -K_2\sqrt{\frac{\epsilon'^3}{\rho}}$ by assumption, the algorithm will not terminate early at this iteration. Since the function f is bounded below by f^*, the event $m_t(\mathbf{x}_t + \mathbf{\Delta}_t) - m_t(\mathbf{x}_t) \leq -K_2\sqrt{\frac{\epsilon'^3}{\rho}}$ which implies $f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -K_{\text{prog}}\sqrt{\frac{\epsilon'^3}{\rho}}$ can happen at most $T_{\text{outer}} = \lceil \frac{\sqrt{\rho}(f(x_0)-f^*)}{K_{\text{prog}}\epsilon'^{1.5}} \rceil$ times.

Thus in the T_{outer} iterations of Algorithm 1 it must be the case that there is at least one iteration T, for which

$$m_T(\mathbf{x}_T + \mathbf{\Delta}_T) - m_T(\mathbf{x}_T) \geq -K_2\sqrt{\frac{\epsilon'^3}{\rho}}.$$

By the definition of the Cubic-Subsolver procedure and assumption that $-K_1 \leq -K_2$, it must be the case at iteration T the routine Cubic-Subsolver used **Case 2**. Now by appealing to Claim 1 and Lemma 7 we must have that $\|\mathbf{\Delta}_T^\star\| \leq \frac{1}{2}\sqrt{\frac{\epsilon'}{\rho}}$ and that $\mathbf{x}_T + \mathbf{\Delta}_T^\star$ is an ϵ'-second-order stationary point of f. As we can see in Line 7 of Algorithm 1, at iteration T the "if statement" will be true. Hence Algorithm 1 will run the final gradient descent loop (Algorithm 2) at iteration T, return the final point and proceed to exit via the break statement. Since the hypotheses of Lemma 8 are satisfied[3] at iteration T, Algorithm 2 will return a final point that is an ϵ-second-order stationary point of f as desired. We can verify the global constant $c = \min\{\frac{K_{\text{prog}}}{8}, c_1, c_2\}$ satisfies the conditions of the theorem.

Remark 4. We can also now do a careful count of the complexity of Algorithm 1. First, note at each outer iteration of Algorithm 1 we require $n_1 \geq \max\left(\frac{M_1}{c_1\epsilon}, \frac{\sigma_1^2}{c_1^2\epsilon^2}\right)\frac{8}{3}\log\frac{2d}{\delta'}$ samples to approximate the gradient and and $n_2 \geq \max(\frac{M_2}{c_2\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{c_2^2\rho\epsilon})\frac{8}{3}\log\frac{2d}{\delta'}$ to approximate the Hessian. The union bound stipulates we should take $\delta'(\epsilon) = \frac{\delta}{3T_{\text{outer}}}$ to control the total failure probability of Algorithm 1. Then as we can see in the Proof of Theorem 1, Algorithm 1 will terminate in at most

$$T_{\text{outer}} = \lceil \frac{8K_{\text{prog}}\sqrt{\rho}(f(x_0) - f^*)}{\epsilon^{3/2}} \rceil \tag{26}$$

iterations. The inner iteration complexity of the Cubic-Subsolver routine is $\mathcal{T}(\epsilon)$. The routine only requires computing the gradient vector once, but recomputes Hessian-vector products at each iteration.

So the gradient complexity becomes

$$\begin{aligned} T_{\text{G}} &\lesssim \mathcal{T}(\epsilon) \times \frac{\sqrt{\rho}(f(x_0) - f^*)}{\epsilon^{1.5}} \times \max\left(\frac{M_1}{c_1\epsilon}, \frac{\sigma_1^2}{c_1^2\epsilon^2}\right)\frac{8}{3}\log\frac{2d}{\delta'} \\ &\sim \tilde{\mathcal{O}}\left(\frac{\sqrt{\rho}\sigma_1^2(f(x_0) - f^*)}{\epsilon^{3.5}}\right) \text{ for } \epsilon \leq \frac{\sigma_1^2}{c_1 M_1}. \end{aligned}$$

Note that $\tilde{O}$ hides logarithmic factors since $\delta'(\epsilon) = \frac{\delta}{3T_{\text{outer}}}$.

The total complexity of Hessian-vector product evaluations is:

$$\begin{aligned} T_{\text{HV}} &\lesssim \frac{K_{\text{prog}}\sqrt{\rho}(f(x_0) - f^*)}{\epsilon^{1.5}} \times \max(\frac{M_2}{c_2\sqrt{\rho\epsilon}}, \frac{\sigma_2^2}{c_2^2\rho\epsilon})\frac{8}{3}\log\frac{2d}{\delta'} \\ &\sim \tilde{\mathcal{O}}\left(\mathcal{T}(\epsilon)\frac{\sigma_2^2(f(x_0) - f^*)}{\sqrt{\rho}\epsilon^3}\right) \text{ for } \epsilon \leq \frac{\sigma_2^4}{c_2^2 M_2^2 \rho}. \end{aligned}$$

Finally, recall the proof of Lemma 8 which shows total complexity of the final gradient descent loop, in Algorithm 2, will be subleading in overall gradient and Hessian-vector product complexity. As before, we can verify the global constant $c = \min\{\frac{K_{\text{prog}}}{8}, c_1, c_2\}$ satisfies the conditions of the Theorem 1.

□

[3] We can also see with this rescaled ϵ' the step-size requirement in Lemma 8 will be satisfied.