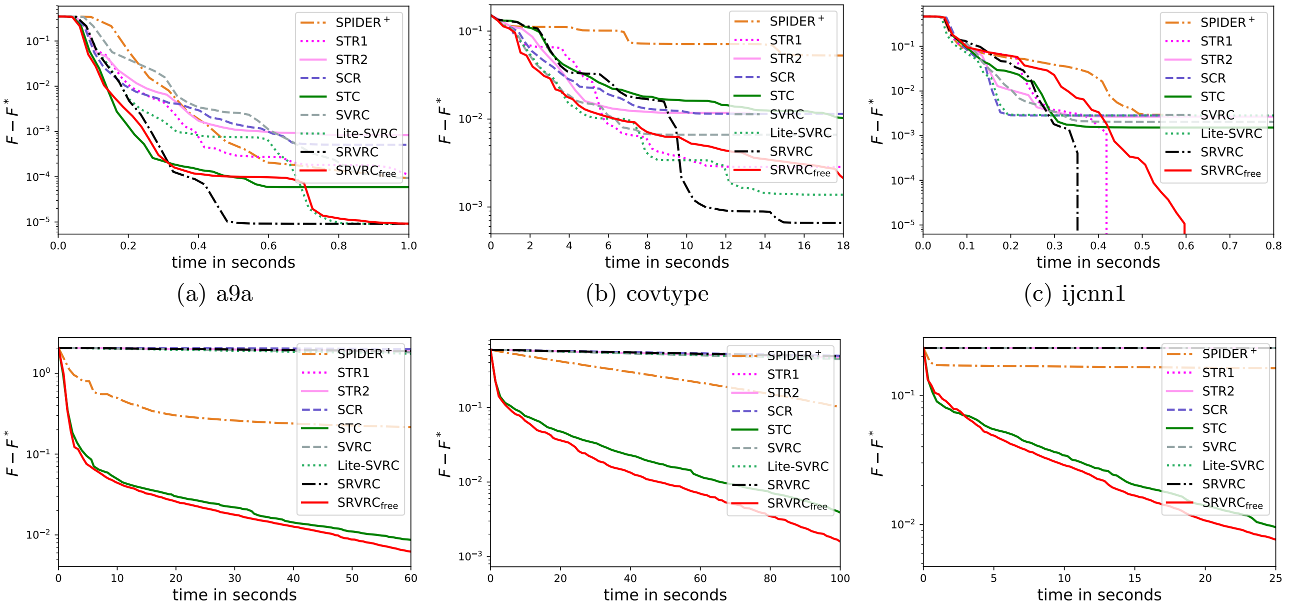
A Experiments



(a) a9a

(b) covtype

(c) ijcnn1

(d) mnist

(e) cifar10

(f) SVHN

Figure 1: Plots of logarithmic function value gap with respect to CPU time (in seconds) for nonconvex regularized binary logistic regression on (a) *a9a* (b) *ovtype* (c) *ijcnn1* and for nonconvex regularized multiclass logistic regression on (d) *mnist* (e) *cifar10* (f) *SVHN*. Best viewed in color.

In this section, we present numerical experiments on different nonconvex Empirical Risk Minimization (ERM) problems and on different datasets to validate the advantage of our proposed SRVRC and SRVRC$_{\text{free}}$ algorithms for finding approximate local minima.

Baselines: We compare our algorithms with the following algorithms: SPIDER+ (Fang et al., 2018), which is the local minimum finding version of SPIDER, stochastic trust region (STR1, STR2) (Shen et al., 2019), subsampled cubic regularization (SCR) (Kohler and Lucchi, 2017), stochastic cubic regularization (STC) (Tripuraneni et al., 2018), stochastic variance-reduced cubic regularization (SVRC) (Zhou et al., 2018d), sample efficient SVRC (Lite-SVRC) (Zhou et al., 2018b; Wang et al., 2018b; Zhang et al., 2018a).

Parameter Settings and Subproblem Solver For each algorithm, we set the cubic penalty parameter M_t adaptively based on how well the model approximates the real objective as suggested in (Cartis et al., 2011a,b; Kohler and Lucchi, 2017). For SRVRC, we set $S^{(g)} = S^{(h)}$ for the simplicity and set gradient and Hessian batch sizes $B_t^{(g)}$ and $B_t^{(h)}$ as follows:

$$\begin{aligned} B_t^{(g)} &= B^{(g)}, B_t^{(h)} = B^{(h)}, && \mathrm{mod}(t, S) = 0, \\ B_t^{(g)} &= \lfloor B^{(g)}/S \rfloor, B_t^{(h)} = \lfloor B^{(h)}/S \rfloor, && \mathrm{mod}(t, S) \neq 0. \end{aligned}$$

For SRVRC$_{\text{free}}$, we set gradient batch sizes $B_t^{(g)}$ the same as SRVRC and Hessian batch sizes $B_t^{(h)} = B^{(h)}$. We tune S over the grid $\{5, 10, 20, 50\}$, $B^{(g)}$ over the grid $\{n, n/10, n/20, n/100\}$, and $B^{(h)}$ over the grid $\{50, 100, 500, 1000\}$ for the best performance. For SCR, SVRC, Lite-SVRC, and SRVRC, we solve the cubic subproblem using the cubic subproblem solver discussed in (Nesterov and Polyak, 2006). For STR1 and STR2, we solve the trust-region subproblem using the exact trust-region subproblem solver discussed in (Conn et al., 2000). For STC and SRVRC$_{\text{free}}$, we use Cubic-Subsolver (Algorithm 3 in Appendix H) to approximately solve the cubic subproblem. All algorithms are carefully tuned for a fair comparison.

Datasets and Optimization Problems We use 6 datasets *a9a*, *covtype*, *ijcnn1* , *mnist*, *cifar10* and *SVHN* from Chang and Lin (2011) . For binary logistic regression problem with a nonconvex regularizer on *a9a*, *covtype*, and *ijcnn1*, we are given training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ are feature vector and output label corresponding to the i-th training example. The nonconvex penalized binary logistic regression is formulated as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n y_i \log\phi(\mathbf{x}_i^\top\mathbf{w}) + (1-y_i)\log[1-\phi(\mathbf{x}_i^\top\mathbf{w})] + \lambda\sum_{i=1}^d \frac{w_i^2}{1+w_i^2},$$

where $\phi(x)$ is the sigmoid function and $\lambda = 10^{-3}$. For multiclass logistic regression problem with a nonconvex regularizer on *mnist*, *cifar10* and *SVHN*, we are given training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}^m$ are feature vectors and multilabels corresponding to the i-th data points. The nonconvex penalized multiclass logistic regression is formulated as follows

$$\min_{\mathbf{W}\in\mathbb{R}^{m\times d}} -\sum_{i=1}^n \frac{1}{n}\langle \mathbf{y}_i, \log[\text{softmax}(\mathbf{W}\mathbf{x}_i)]\rangle + \lambda\sum_{i=1}^m\sum_{j=1}^d 1 + w_{i,j}^2,$$

where $\text{softmax}(\mathbf{a}) = \exp(\mathbf{a})/\sum_{i=1}^d \exp(a_i)$ is the softmax function and $\lambda = 10^{-3}$.

We plot the logarithmic function value gap with respect to CPU time in Figure 1. From Figure 1(a) to 1(f), we can see that for the low dimension optimization task on *a9a*, *covtype* and *ijcnn1*, our SRVRC outperforms all the other algorithms with respect to CPU time. We can also observe that the stochastic trust region method STR1 is better than STR2, which is well-aligned with our discussion before. The SPIDER+ does not perform as well as other second-order methods, even though its stochastic gradient and Hessian complexity is comparable to second-order methods in theory. Meanwhile, we also notice that $\text{SRVRC}_{\text{free}}$ always outperforms STC, which suggests that the variance reduction technique is useful. For high dimension optimization task *mnist*, *cifar10* and *SVHN*, only SPIDER+, STC and $\text{SRVRC}_{\text{free}}$ are able to make notable progress and $\text{SRVRC}_{\text{free}}$ outperforms the other two. This is again consistent with our theory and discussions in Section 5. Overall, our experiments clearly validate the advantage of SRVRC and $\text{SRVRC}_{\text{free}}$, and corroborate the theory of both algorithms.

B Proofs in Section 4

We define the filtration $\mathcal{F}_t = \sigma(\mathbf{x}_0, ..., \mathbf{x}_t)$ as the σ-algebra of $\mathbf{x}_0$ to $\mathbf{x}_t$. Recall that $\mathbf{v}_t$ and $\mathbf{U}_t$ are the semi-stochastic gradient and Hessian respectively, $\mathbf{h}_t$ is the update parameter, and M_t is the cubic penalty parameter appeared in Algorithm 1 and Algorithm 2. We denote $m_t(\mathbf{h}) := \mathbf{v}^\top\mathbf{h} + \mathbf{h}^\top\mathbf{U}_t\mathbf{h}/2 + M_t\|\mathbf{h}\|_2^3/6$ and $\mathbf{h}_t^* = \text{argmin}_{\mathbf{h}\in\mathbb{R}^d} m_t(\mathbf{h})$. In this section, we define $\delta = \xi/(2T)$ for the simplicity.

B.1 Proof of Theorem 4.2

To prove Theorem 4.2, we need the following lemma adapted from Zhou et al. (2018d), which characterizes that $\mu(\mathbf{x}_t + \mathbf{h})$ can be bounded by $\|\mathbf{h}\|_2$ and the norm of difference between semi-stochastic gradient and Hessian.

Lemma B.1. Suppose that $m_t(\mathbf{h}) := \mathbf{v}_t^\top\mathbf{h} + \mathbf{h}^\top\mathbf{U}_t\mathbf{h}/2 + M_t\|\mathbf{h}\|_2^3/6$ and $\mathbf{h}_t^* = \text{argmin}_{\mathbf{h}\in\mathbb{R}^d} m_t(\mathbf{h})$. If $M_t/\rho \geq 2$, then for any $\mathbf{h} \in \mathbb{R}^d$, we have

$$\mu(\mathbf{x}_t + \mathbf{h}) \leq 9\Big[M_t^3\rho^{-3/2}\|\mathbf{h}\|_2^3 + M_t^{3/2}\rho^{-3/2}\big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2^{3/2} + \rho^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3 \\ + M_t^{3/2}\rho^{-3/2}\|\nabla m_t(\mathbf{h})\|_2^{3/2} + M_t^3\rho^{-3/2}\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^*\|_2\big|^3\Big].$$

Next lemma gives bounds on the inner products $\langle\nabla F(\mathbf{x}_t) - \mathbf{v}_t, \mathbf{h}\rangle$ and $\langle\big(\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big)\mathbf{h}, \mathbf{h}\rangle$.

Lemma B.2. For any $\mathbf{h} \in \mathbb{R}^d$, we have

$$\langle\nabla F(\mathbf{x}_t) - \mathbf{v}_t, \mathbf{h}\rangle \leq \frac{\rho}{8}\|\mathbf{h}\|_2^3 + \frac{6\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2}}{5\sqrt{\rho}},$$

$$\big\langle\big(\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big)\mathbf{h}, \mathbf{h}\big\rangle \leq \frac{\rho}{8}\|\mathbf{h}\|_2^3 + \frac{10}{\rho^2}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3.$$

We also need the following two lemmas, which show that semi-stochastic gradient and Hessian $\mathbf{v}_t$ and $\mathbf{U}_t$ estimators are good approximations to true gradient and Hessian.

Lemma B.3. Suppose that $\{B_k^{(g)}\}$ satisfies (4.1) and (4.3), then conditioned on $\mathcal{F}_{\lfloor t/S^{(g)}\rfloor\cdot S^{(g)}}$, with probability at least $1 - \delta\cdot(t - \lfloor t/S^{(g)}\rfloor\cdot S^{(g)})$, we have that for all $\lfloor t/S^{(g)}\rfloor\cdot S^{(g)} \leq k \leq t$,

$$\|\nabla F(\mathbf{x}_k) - \mathbf{v}_k\|_2^2 \leq \frac{\epsilon^2}{30}. \tag{B.1}$$

Lemma B.4. Suppose that $\{B_k^{(h)}\}$ satisfies (4.2) and (4.4), then conditioned on $\mathcal{F}_{\lfloor t/S^{(h)}\rfloor \cdot S^{(h)}}$, with probability at least $1 - \delta \cdot (t - \lfloor t/S^{(h)}\rfloor \cdot S^{(h)})$, we have that for all $\lfloor t/S^{(h)}\rfloor \cdot S^{(h)} \le k \le t$,

$$\|\nabla^2 F(\mathbf{x}_k) - \mathbf{U}_k\|_2^2 \le \frac{\rho\epsilon}{20}. \tag{B.2}$$

Given all the above lemmas, we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. Suppose that SRVRC terminates at iteration $T^* - 1$, then $\|\mathbf{h}_t\|_2 > \sqrt{\epsilon/\rho}$ for all $0 \le t \le T^* - 1$. We have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\le F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbf{h}_t\rangle + \frac{1}{2}\langle \mathbf{h}_t, \nabla^2 F(\mathbf{x}_t)\mathbf{h}_t\rangle + \frac{\rho}{6}\|\mathbf{h}_t\|_2^3 \\
&= F(\mathbf{x}_t) + m_t(\mathbf{h}_t) + \frac{\rho - M_t}{6}\|\mathbf{h}_t\|_2^3 + \langle \mathbf{h}_t, \nabla F(\mathbf{x}_t) - \mathbf{v}_t\rangle + \frac{1}{2}\langle \mathbf{h}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t)\mathbf{h}_t\rangle \\
&\le F(\mathbf{x}_t) - \frac{\rho}{2}\|\mathbf{h}_t\|_2^3 + \frac{\rho}{4}\|\mathbf{h}_t\|_2^3 + \frac{6\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2}}{5\sqrt{\rho}} + \frac{10}{\rho^2}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|_2^3 \\
&= F(\mathbf{x}_t) - \frac{\rho}{4}\|\mathbf{h}_t\|_2^3 + \frac{6\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2}}{5\sqrt{\rho}} + \frac{10}{\rho^2}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|_2^3, \qquad \text{(B.3)}
\end{aligned}$$

where the second inequality holds due to the fact that $m_t(\mathbf{h}_t) \le m_t(\mathbf{0}) = 0$, $M_t = 4\rho$ and Lemma B.2. By Lemmas B.3 and B.4, with probability at least $1 - 2T\delta$, for all $0 \le t \le T - 1$, we have that

$$\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2} \le \frac{\epsilon^{3/2}}{12}, \quad \|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|_2^3 \le \frac{(\rho\epsilon)^{3/2}}{80} \tag{B.4}$$

for all $0 \le t \le T - 1$. Substituting (B.4) into (B.3), we have

$$F(\mathbf{x}_{t+1}) \le F(\mathbf{x}_t) - \frac{\rho}{4}\|\mathbf{h}_t\|_2^3 + \frac{9\rho^{-1/2}\epsilon^{3/2}}{40}. \tag{B.5}$$

Telescoping (B.5) from $t = 0, \ldots, T^* - 1$, we have

$$\Delta_F \ge F(\mathbf{x}_0) - F(\mathbf{x}_{T^*}) \ge \rho \cdot T^* \cdot (\epsilon/\rho)^{3/2}/4 - 9/40 \cdot \rho^{-1/2}\epsilon^{3/2} \cdot T^* = \rho^{-1/2}\epsilon^{3/2} \cdot T^*/40. \tag{B.6}$$

Recall that we have $T \ge 40\Delta_F\sqrt{\rho}/\epsilon^{3/2}$ from the condition of Theorem 4.2, then by (B.6), we have $T^* \le T$. Thus, we have $\|\mathbf{h}_{T^*-1}\|_2 \le \sqrt{\epsilon/\rho}$. Denote $\widetilde{T} = T^* - 1$, then we have

$$\begin{aligned}
\mu(\mathbf{x}_{\widetilde{T}+1}) &= \mu(\mathbf{x}_{\widetilde{T}} + \mathbf{h}_{\widetilde{T}}) \\
&\le 9\Big[M_{\widetilde{T}}^3\rho^{-3/2}\|\mathbf{h}_{\widetilde{T}}\|_2^3 + M_{\widetilde{T}}^{3/2}\rho^{-3/2}\big\|\nabla F(\mathbf{x}_{\widetilde{T}}^s) - \mathbf{v}_{\widetilde{T}}\big\|_2^{3/2} + \rho^{-3/2}\big\|\nabla^2 F(\mathbf{x}_{\widetilde{T}}) - \mathbf{U}_{\widetilde{T}}\big\|_2^3\Big] \\
&\le 600\epsilon^{3/2},
\end{aligned}$$

where the first inequality holds due to Lemma B.1 with $\nabla m_{\widetilde{T}}(\mathbf{h}_{\widetilde{T}}) = 0$ and $\|\mathbf{h}_{\widetilde{T}}\|_2 = \|\mathbf{h}_{\widetilde{T}}^*\|_2$. This completes our proof. □

B.2 Proof of Corollary 4.3

Proof of Corollary 4.3. Suppose that SRVRC terminates at $T^* - 1 \le T - 1$ iteration. Telescoping (B.5) from $t = 0$ to $T^* - 1$, we have

$$\Delta_F \ge F(\mathbf{x}_0) - F(\mathbf{x}_{T^*}) \ge \rho \sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^3/4 - 9\rho^{-1/2}\epsilon^{3/2}/40 \cdot T = \rho \sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^3/4 - 9 \cdot \Delta_F, \tag{B.7}$$

where the last inequality holds since T is set to be $40\Delta_F\sqrt{\rho}/\epsilon^{3/2}$ as the conditions of Corollary 4.3 suggests. (B.7) implies that $\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^3 \le 40\Delta_F/\rho$. Thus, we have

$$\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^2 \le (T^*)^{1/3}\bigg(\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^3\bigg)^{2/3} \le \bigg(\frac{40\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg)^{1/3} \cdot \bigg(\frac{40\Delta_F}{\rho}\bigg)^{2/3} = \frac{40\Delta_F}{\rho^{1/2}\epsilon^{1/2}}, \tag{B.8}$$

where the first inequality holds due to Hölder's inequality inequality, and the second inequality is due to $T^* \le T = 40\Delta_F\sqrt{\rho}/\epsilon^{3/2}$. We first consider the total gradient sample complexity $\sum_{t=0}^{T^*-1} B_t^{(g)}$, which can be bounded as

$$\begin{aligned}
&\sum_{t=0}^{T^*-1} B_t^{(g)}\\
&= \sum_{\mathrm{mod}(t,S^{(g)})=0} B_t^{(g)} + \sum_{\mathrm{mod}(t,S^{(g)})\neq 0} B_t^{(g)}\\
&= \sum_{\mathrm{mod}(t,S^{(g)})=0} \min\bigg\{n, 1440\frac{M^2\log^2(d/\delta)}{\epsilon^2}\bigg\} + \sum_{\mathrm{mod}(t,S^{(g)})\neq 0} \min\bigg\{n, 1440L^2\log^2(d/\delta)\frac{S^{(g)}\|\mathbf{h}_{t-1}\|_2^2}{\epsilon^2}\bigg\}\\
&\le C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{T^*}{S^{(g)}}\bigg(n\wedge\frac{M^2}{\epsilon^2}\bigg) + \bigg(\frac{L^2S^{(g)}}{\epsilon^2}\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^2\bigg)\wedge nT^*\bigg]\\
&\le 40C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}S^{(g)}}\bigg(n\wedge\frac{M^2}{\epsilon^2}\bigg) + \bigg(\frac{\Delta_F L^2 S^{(g)}}{\rho^{1/2}\epsilon^{5/2}}\bigg)\wedge\frac{n\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg]\\
&= \widetilde{O}\bigg(n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F}{\epsilon^{3/2}}\bigg[\sqrt{\rho}n\wedge\frac{L\sqrt{n}}{\sqrt{\epsilon}}\wedge\frac{LM}{\epsilon^{3/2}}\bigg]\bigg),
\end{aligned}$$

where $C_1 = 1440\log^2(d/\delta)$, the second inequality holds due to (B.8), and the last equality holds due to the choice of $S^{(g)} = \sqrt{\rho\epsilon}/L\cdot\sqrt{n\wedge M^2/\epsilon^2}$. We then consider the total Hessian sample complexity $\sum_{t=0}^{T^*-1} B_t^{(h)}$, which can be bounded as

$$\begin{aligned}
&\sum_{t=0}^{T^*-1} B_t^{(h)}\\
&= \sum_{\mathrm{mod}(t,S^{(h)})=0} B_t^{(h)} + \sum_{\mathrm{mod}(t,S^{(h)})\neq 0} B_t^{(h)}\\
&= \sum_{\mathrm{mod}(t,S^{(h)})=0} \min\bigg\{n, 800\frac{L^2\log^2(d/\delta)}{\rho\epsilon}\bigg\} + \sum_{\mathrm{mod}(t,S^{(h)})\neq 0} \min\bigg\{n, 800\rho\log^2(d/\delta)\frac{S^{(h)}\|\mathbf{h}_{t-1}\|_2^2}{\epsilon}\bigg\}\\
&\le C_2\bigg[n\wedge\frac{L^2}{\rho\epsilon} + \frac{T^*}{S^{(h)}}\bigg(n\wedge\frac{L^2}{\rho\epsilon}\bigg) + \frac{\rho S^{(h)}}{\epsilon}\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^2\bigg]\\
&\le 40C_2\bigg[n\wedge\frac{L^2}{\rho\epsilon} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}S^{(h)}}\bigg(n\wedge\frac{L^2}{\rho\epsilon}\bigg) + \frac{\Delta_F\rho^{1/2}S^{(h)}}{\epsilon^{3/2}}\bigg]\\
&= \widetilde{O}\bigg[n\wedge\frac{L^2}{\rho\epsilon} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\sqrt{n\wedge\frac{L^2}{\rho\epsilon}}\bigg],
\end{aligned}$$

where $C_2 = 800\log^2(d/\delta)$, the second inequality holds due to (B.8), and the last equality holds due to the choice of $S^{(h)} = \sqrt{n\wedge L/(\rho\epsilon)}$. □

C Proofs in Section 5

In this section, we denote $\delta = \xi/(3T)$ for simplicity.

C.1 Proof of Theorem 5.1

We need the following two lemmas, which bound the variance of semi-stochastic gradient and Hessian estimators.

Lemma C.1. Suppose that $\{B_k^{(g)}\}$ satisfies (5.2) and (5.3), then conditioned on $\mathcal{F}_{\lfloor t/S\rfloor\cdot S}$, with probability at least $1-\delta\cdot(t-\lfloor t/S\rfloor\cdot S)$, we have that for all $\lfloor t/S\rfloor\cdot S\le k\le t$,

$$\|\nabla F(\mathbf{x}_k)-\mathbf{v}_k\|_2^2\le\frac{\epsilon^2}{55}.$$

Proof of Lemma C.1. The proof is very similar to that of Lemma B.3, hence we omit it. □

Lemma C.2. Suppose that $\{B_k^{(h)}\}$ satisfies (5.1), then conditioned on $\mathcal{F}_k$, with probability at least $1-\delta$, we have that

$$\|\nabla^2 F(\mathbf{x}_k) - \mathbf{U}_k\|_2^2 \leq \frac{\rho\epsilon}{30}.$$

Proof of Lemma C.2. The proof is very similar to that of Lemma B.4, hence we omit it. □

We have the following lemma to guarantee that by Algorithm 3 Cubic-Subsolver, the output $\mathbf{h}_t$ satisfies that sufficient decrease of function value will be made and the total number of iterations is bounded by T'.

Lemma C.3. For any $t \geq 0$, suppose that $\|\mathbf{h}_t^*\|_2 \geq \sqrt{\epsilon/\rho}$ or $\|\mathbf{v}_t\|_2 \geq \max\{M_t\epsilon/(2\rho), \sqrt{LM_t/2}(\epsilon/\rho)^{3/4}\}$. We set $\eta = 1/(16L)$. Then for $\epsilon < 16L^2\rho/M_t^2$, with probability at least $1-\delta$, Cubic-Subsolver$(\mathbf{U}_t, \mathbf{v}_t, M_t, \eta, \sqrt{\epsilon/\rho}, 0.5, \delta)$ will return $\mathbf{h}_t$ satisfying $m_t(\mathbf{h}_t) \leq -M_t\rho^{-3/2}\epsilon^{3/2}/24$. within

$$T' = C_S\frac{L}{M_t\sqrt{\epsilon/\rho}}$$

iterations, where $C_S > 0$ is a constant.

We have the following lemma which provides the guarantee for the dynamic of gradient steps in Cubic-Finalsolver.

Lemma C.4. (Carmon and Duchi, 2016) For $\mathbf{b}, \mathbf{A}, \tau$, suppose that $\|\mathbf{A}\|_2 \leq L$. We denote that $g(\mathbf{h}) = \mathbf{b}^\top\mathbf{h} + \mathbf{h}^\top\mathbf{A}\mathbf{h}/2 + \tau/6 \cdot \|\mathbf{h}\|_2^3$, $\mathbf{s} = \mathrm{argmin}_{\mathbf{h}\in\mathbb{R}^d}\, g(\mathbf{h})$, and let R be

$$R = \frac{L}{2\tau} + \sqrt{\left(\frac{L}{2\tau}\right)^2 + \frac{\|\mathbf{b}\|_2}{\tau}}.$$

Then for Cubic-Finalsolver, suppose that $\eta < (4(L+\tau R))^{-1}$, then each iterate Δ in Cubic-Finalsolver satisfies that $\|\Delta\|_2 \leq \|\mathbf{s}\|_2$, and $g(\mathbf{h})$ is $(L+2\tau R)$-smooth.

With these lemmas, we begin our proof of Theorem 5.1.

Proof of Theorem 5.1. Suppose that SRVRC$_{\text{free}}$ terminates at iteration T^*-1. Then $T^* \leq T$. We first claim that $T^* < T$. Otherwise, suppose $T^* = T$, then we have that for all $0 \leq t < T^*$,

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbf{h}_t\rangle + \frac{1}{2}\langle \mathbf{h}_t, \nabla^2 F(\mathbf{x}_t)\mathbf{h}_t\rangle + \frac{\rho}{6}\|\mathbf{h}_t\|_2^3 \\
&= F(\mathbf{x}_t) + m_t(\mathbf{h}_t) + \frac{\rho - M_t}{6}\|\mathbf{h}_t\|_2^3 + \langle \mathbf{h}_t, \nabla F(\mathbf{x}_t) - \mathbf{v}_t\rangle + \frac{1}{2}\langle \mathbf{h}_t, (\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t)\mathbf{h}_t\rangle \\
&\leq F(\mathbf{x}_t) - \frac{\rho}{4}\|\mathbf{h}_t\|_2^3 + m_t(\mathbf{h}_t) + \frac{6\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2}}{5\sqrt{\rho}} + \frac{10}{\rho^2}\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|_2^3,
\end{aligned} \tag{C.1}$$

where the second inequality holds due to $M_t = 4\rho$ and Lemma B.2. By Lemma C.3 and union bound, we know that with probability at least $1 - T\delta$, we have

$$m_t(\mathbf{h}_t) \leq -M_t\rho^{-3/2}\epsilon^{3/2}/24 = -\rho^{-1/2}\epsilon^{3/2}/6, \tag{C.2}$$

where we use the fact that $M_t = 4\rho$. By Lemmas C.1 and C.2, we know that with probability at least $1-2T\delta$, for all $0 \leq t \leq T^*-1$, we have

$$\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2} \leq \epsilon^{3/2}/20, \quad \|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\|_2^3 \leq (\rho\epsilon)^{3/2}/160. \tag{C.3}$$

Substituting (C.2) and (C.3) into (C.1), we have

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) \leq -\rho^{-1/2}\epsilon^{3/2}/6 - \rho\|\mathbf{h}_t\|_2^3/4 + \rho^{-1/2}\epsilon^{3/2}/8 \leq -\rho\|\mathbf{h}_t\|_2^3/4 - \rho^{-1/2}\epsilon^{3/2}/24. \tag{C.4}$$

Telescoping (C.4) from $t = 0$ to $T^* - 1$, we have

$$\Delta_F \geq F(\mathbf{x}_0) - F(\mathbf{x}_{T^*}) \geq \rho \sum_{t=0}^{T^*-1} \|\mathbf{h}_t\|_2^3/4 + \rho^{-1/2}\epsilon^{3/2} \cdot T^*/24 > \rho \sum_{t=0}^{T^*-1} \|\mathbf{h}_t\|_2^3/4 + \Delta_F, \tag{C.5}$$

where the last inequality holds since we assume $T^* = T \geq 25\Delta_F \rho^{1/2}\epsilon^{-3/2}$ from the condition of Theorem 5.1. (C.5) leads to a contradiction, thus we have $T^* < T$. Therefore, by union bound, with probability at least $1 - 3T\delta$, Cubic-Finalsolver is executed by SRVRC$_{\text{free}}$ at $T^* - 1$ iteration. We have that $\|\mathbf{v}_{T^*-1}\|_2 < \max\{M_{T^*-1}\epsilon/(2\rho), \sqrt{LM_{T^*-1}/2}(\epsilon/\rho)^{3/4}\}$ and $\|\mathbf{h}^*_{T^*-1}\|_2 < \sqrt{\epsilon/\rho}$ by Lemma C.3.

The only thing left is to check that we indeed find a second-order stationary point, $\mathbf{x}_{T^*}$, by Cubic-Finalsolver. We first need to check that the choice of $\eta = 1/(16L)$ satisfies that $1/\eta > 4(L + M_t R)$ by Lemma C.4, where

$$R = \frac{L}{2M_{T^*-1}} + \sqrt{\left(\frac{L}{2M_{T^*-1}}\right)^2 + \frac{\|\mathbf{v}_{T^*-1}\|_2}{M_{T^*-1}}},$$

We can check that with the assumption that $\|\mathbf{v}_{T^*-1}\|_2 < \max\{M_{T^*-1}\epsilon/(2\rho), \sqrt{LM_{T^*-1}/2}(\epsilon/\rho)^{3/4}\}$, if $\epsilon < 4L^2\rho/M^2_{T^*-1}$, then $1/\eta > 4(L + M_{T^*-1}R)$ holds.

For simplicity, we denote $\widetilde{T} = T^* - 1$. Then we have

$$\begin{aligned}
\mu(\mathbf{x}_{\widetilde{T}} + \mathbf{h}_{\widetilde{T}}) &\leq 9\Big[M^3_{\widetilde{T}}\rho^{-3/2}\|\mathbf{h}_{\widetilde{T}}\|_2^3 + M^{3/2}_{\widetilde{T}}\rho^{-3/2}\big\|\nabla F(\mathbf{x}_{\widetilde{T}}) - \mathbf{v}_{\widetilde{T}}\big\|_2^{3/2} + \rho^{-3/2}\big\|\nabla^2 F(\mathbf{x}_{\widetilde{T}}) - \mathbf{U}_{\widetilde{T}}\big\|_2^3 \\
&\quad + M^{3/2}_{\widetilde{T}}\rho^{-3/2}\|\nabla m_{\widetilde{T}}(\mathbf{h}_{\widetilde{T}})\|_2^{3/2} + M^3_{\widetilde{T}}\rho^{-3/2}\big|\|\mathbf{h}_{\widetilde{T}}\|_2 - \|\mathbf{h}^*_{\widetilde{T}}\|_2\big|^3\Big] \\
&\leq 9\Big[2M^3_{\widetilde{T}}\rho^{-3/2}\|\mathbf{h}^*_{\widetilde{T}}\|_2^3 + M^{3/2}_{\widetilde{T}}\rho^{-3/2}\big\|\nabla F(\mathbf{x}_{\widetilde{T}}) - \mathbf{v}_{\widetilde{T}}\big\|_2^{3/2} + \rho^{-3/2}\big\|\nabla^2 F(\mathbf{x}_{\widetilde{T}}) - \mathbf{U}_{\widetilde{T}}\big\|_2^3 \\
&\quad + M^{3/2}_{\widetilde{T}}\rho^{-3/2}\|\nabla m_{\widetilde{T}}(\mathbf{h}_{\widetilde{T}})\|_2^{3/2}\Big] \\
&\leq 1300\epsilon^{3/2},
\end{aligned}$$

where the first inequality holds due to Lemma B.1, the second inequality holds due to the fact that $\|\mathbf{h}_{\widetilde{T}}\|_2 \leq \|\mathbf{h}^*_{\widetilde{T}}\|_2$ from Lemma C.4, the last inequality holds due to the facts that $\|\nabla m_{\widetilde{T}}(\mathbf{h}_{\widetilde{T}})\|_2 \leq \epsilon$ from Cubic-Finalsolver and $\|\mathbf{h}^*_{\widetilde{T}}\|_2 \leq \sqrt{\epsilon/\rho}$ by Lemma C.3. □

C.2 Proof of Corollary 5.2

We have the following lemma to bound the total number of iterations T'' of Algorithm 4 Cubic-Finalsolver.

Lemma C.5. If $\epsilon < 4L^2\rho/M_t^2$, then Cubic-Finalsolver will terminate within $T'' = C_F L/\sqrt{\rho\epsilon}$ iterations, where $C_F > 0$ is a constant.

Proof of Corollary 5.2. We have that

$$\sum_{t=0}^{T^*-1} \|\mathbf{h}_t\|_2^2 \leq (T^*)^{1/3}\left(\sum_{t=0}^{T^*-1} \|\mathbf{h}_t\|_2^3\right)^{2/3} \leq \left(\frac{25\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\right)^{1/3} \cdot \left(\frac{4\Delta_F}{\rho}\right)^{2/3} \leq \frac{\Delta_F}{8\rho^{1/2}\epsilon^{1/2}}, \tag{C.6}$$

where the first inequality holds due to Hölder's inequality, the second inequality holds due to the facts that $T^* \leq T = 25\Delta_F\rho^{1/2}/\epsilon^{3/2}$ and $\Delta_F \geq \rho\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^3/4$ by (C.5). We first consider the total stochastic gradient computations, $\sum_{t=0}^{T^*-1} B_t^{(g)}$, which can be bounded as

$$\begin{aligned}
&\sum_{t=0}^{T^*-1} B_t^{(g)} \\
&= \sum_{\mathrm{mod}(t,S^{(g)})=0} B_t^{(g)} + \sum_{\mathrm{mod}(t,S^{(g)})\neq 0} B_t^{(g)} \\
&= \sum_{\mathrm{mod}(t,S^{(g)})=0} \min\bigg\{n, 2640\frac{M^2\log^2(d/\delta)}{\epsilon^2}\bigg\} + \sum_{\mathrm{mod}(t,S^{(g)})\neq 0} \min\bigg\{n, 2640L^2\log^2(d/\delta)\frac{S^{(g)}\|\mathbf{h}_{t-1}\|_2^2}{\epsilon^2}\bigg\} \\
&\leq C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{T^*}{S^{(g)}}\bigg(n\wedge\frac{M^2}{\epsilon^2}\bigg) + \bigg(\frac{L^2S^{(g)}}{\epsilon^2}\sum_{t=0}^{T^*-1}\|\mathbf{h}_t\|_2^2\bigg)\wedge nT^*\bigg] \\
&\leq 8C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}S^{(g)}}\bigg(n\wedge\frac{M^2}{\epsilon^2}\bigg) + \bigg(\frac{\Delta_F L^2 S^{(g)}}{\rho^{1/2}\epsilon^{5/2}}\bigg)\wedge\frac{n\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg] \\
&= 8C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg(\frac{1}{S^{(g)}}\bigg(n\wedge\frac{M^2}{\epsilon^2}\bigg) + \bigg(\frac{L^2S^{(g)}}{\rho\epsilon}\bigg)\wedge n\bigg)\bigg] \\
&= 8C_1\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg(n\wedge\frac{L\sqrt{n}}{\sqrt{\rho\epsilon}}\wedge\frac{LM}{\rho^{1/2}\epsilon^{3/2}}\bigg)\bigg],
\end{aligned} \tag{C.7}$$

where $C_1 = 2640\log^2(d/\delta)$, the second inequality holds due to (C.6), the last equality holds due to the fact $S^{(g)} = \sqrt{\rho\epsilon}/L\cdot\sqrt{n\wedge M^2/\epsilon^2}$. We now consider the total amount of Hessian-vector product computations $\mathcal{T}$, which includes $\mathcal{T}_1$ from Cubic-Subsolver and $\mathcal{T}_2$ from Cubic-Finalsolver. By Lemma C.3, we know that at k-th iteration of SRVRC$_{\text{free}}$, Cubic-Subsolver has T' iterations, which needs $B_k^{(h)}$ Hessian-vector product computations each time. Thus, we have

$$\mathcal{T}_1 = \sum_{k=0}^{T^*-1} T'\cdot B_k^{(h)} \leq C_2\bigg(T\cdot T'\cdot\bigg[n\wedge\frac{L^2}{\rho\epsilon}\bigg]\bigg) \leq 25C_2\bigg(T'\frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg[n\wedge\frac{L^2}{\rho\epsilon}\bigg]\bigg) \leq 7C_2C_S\bigg(\frac{L\Delta_F}{\epsilon^2}\cdot\bigg[n\wedge\frac{L^2}{\rho\epsilon}\bigg]\bigg), \tag{C.8}$$

where $C_2 = 1200\log^2(d/\delta)$, the first inequality holds due to the fact that $B_k^{(h)} = C_2 n\wedge(L^2/\rho\epsilon)$, the second inequality holds due to the fact that $T = 25\Delta_F\rho^{1/2}/\epsilon^{3/2}$, the last inequality holds due to the fact that $T' = C_S L/M_t\cdot\sqrt{\rho/\epsilon} = C_S L/(4\sqrt{\rho\epsilon})$. For $\mathcal{T}_2$, we have

$$\mathcal{T}_2 = B_{T^*-1}^{(h)}\cdot T'' \leq C_2T''\bigg[n\wedge\frac{L^2}{\rho\epsilon}\bigg] \leq C_2C_F\bigg(\frac{L}{\sqrt{\rho\epsilon}}\cdot\bigg[n\wedge\frac{L^2}{\rho\epsilon}\bigg]\bigg), \tag{C.9}$$

where the first inequality holds due to the fact that $B_{T^*-1}^{(h)} = C_2 n\wedge(L^2/\rho\epsilon)$, the second inequality holds due to the fact that $T'' = C_F L/\sqrt{\rho\epsilon}$ by Lemma C.5. Since at each iteration we need $B_{T^*-1}^{(h)}$ Hessian-vector computations. Combining (C.7), (C.8) and (C.9), we know that the total stochastic gradient and Hessian-vector product computations are bounded as

$$\begin{aligned}
&\sum_{t=0}^{T^*-1} B_t^{(g)} + \mathcal{T}_1 + \mathcal{T}_2 \\
&= \widetilde{O}\bigg[n\wedge\frac{M^2}{\epsilon^2} + \frac{\Delta_F\rho^{1/2}}{\epsilon^{3/2}}\bigg(n\wedge\frac{L\sqrt{n}}{\sqrt{\rho\epsilon}}\wedge\frac{LM}{\rho^{1/2}\epsilon^{3/2}}\bigg) + \bigg(\frac{L\Delta_F}{\epsilon^2} + \frac{L}{\sqrt{\rho\epsilon}}\bigg)\cdot\bigg(n\wedge\frac{L^2}{\rho\epsilon}\bigg)\bigg].
\end{aligned} \tag{C.10}$$

□

D Proofs of Technical Lemmas in Appendix B

D.1 Proof of Lemma B.1

We have the following lemmas from Zhou et al. (2018d)

Lemma D.1. (Zhou et al., 2018d) If $M_t \geq 2\rho$, then we have

$$\|\nabla F(\mathbf{x}_t + \mathbf{h})\|_2 \leq M_t\|\mathbf{h}\|_2^2 + \big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2 + \frac{1}{M_t}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^2 + \|\nabla m_t(\mathbf{h})\|_2.$$

Lemma D.2. (Zhou et al., 2018d) If $M_t \geq 2\rho$, then we have

$$-\lambda_{\min}(\nabla^2 F(\mathbf{x}_t + \mathbf{h})) \leq M_t\|\mathbf{h}\|_2 + \big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2 + M_t\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^*\|_2\big|.$$

Proof of Lemma B.1. By Lemma D.1, we have

$$\begin{aligned}\|\nabla F(\mathbf{x}_t + \mathbf{h})\|_2^{3/2} &\leq \Big[M_t\|\mathbf{h}\|_2^2 + \big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2 + \frac{1}{M_t}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^2 + \|\nabla m_t(\mathbf{h})\|_2\Big]^{3/2}\\ &\leq 2\Big[M_t^{3/2}\|\mathbf{h}\|_2^3 + \big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2^{3/2} + M_t^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3 + \|\nabla m_t(\mathbf{h})\|_2^{3/2}\Big], \end{aligned}\tag{D.1}$$

where the second inequality holds due to the fact that for any $a, b, c \geq 0$, we have $(a+b+c)^{3/2} \leq 2(a^{3/2}+b^{3/2}+c^{3/2})$. By Lemma D.2, we have

$$\begin{aligned}-\rho^{-3/2}\lambda_{\min}(\nabla^2 F(\mathbf{x}_t + \mathbf{h}))^3 &\leq \rho^{-3/2}\Big[M_t\|\mathbf{h}\|_2 + \big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2 + M_t\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^*\|_2\big|\Big]^3\\ &\leq 9\rho^{-3/2}\Big[M_t^3\|\mathbf{h}\|_2^3 + \big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3 + M_t^3\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^*\|_2\big|^3\Big], \end{aligned}\tag{D.2}$$

where the second inequality holds due to the fact that for any $a, b, c \geq 0$, we have $(a + b + c)^3 \leq 9(a^3 + b^3 + c^3)$. Thus we have

$$\begin{aligned}\mu(\mathbf{x}_t + \mathbf{h}) &= \max\{\|\nabla F(\mathbf{x}_t + \mathbf{h})\|_2^{3/2}, -\rho^{-3/2}\lambda_{\min}(\nabla^2 F(\mathbf{x}_t + \mathbf{h}))^3\}\\ &\leq 9\Big[M_t^3\rho^{-3/2}\|\mathbf{h}\|_2^3 + M_t^{3/2}\rho^{-3/2}\big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2^{3/2} + \rho^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3\\ &\quad + M_t^{3/2}\rho^{-3/2}\|\nabla m_t(\mathbf{h})\|_2^{3/2} + M_t^3\rho^{-3/2}\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^*\|_2\big|^3\Big],\end{aligned}$$

where the inequality holds due to (D.1), (D.2) and the fact that $M_t \geq 4\rho$. □

D.2 Proof of Lemma B.2

Proof of Lemma B.2. We have

$$\langle\nabla F(\mathbf{x}_t) - \mathbf{v}_t, \mathbf{h}\rangle \leq \big\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\big\|_2\|\mathbf{h}\|_2 \leq \frac{\rho}{8}\|\mathbf{h}\|_2^3 + \frac{6\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|_2^{3/2}}{5\sqrt{\rho}},$$

where the first inequality holds due to CauchySchwarz inequality, the second inequality holds due to Young's inequality. We also have

$$\big\langle\big(\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big)\mathbf{h}, \mathbf{h}\big\rangle \leq \big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2\|\mathbf{h}\|_2^2 \leq \frac{\rho}{8}\|\mathbf{h}\|_2^3 + \frac{10}{\rho^2}\big\|\nabla^2 F(\mathbf{x}_t) - \mathbf{U}_t\big\|_2^3,$$

where the first inequality holds due to CauchySchwarz inequality, the second inequality holds due to Young's inequality. □

D.3 Proof of Lemma B.3

We need the following lemma:

Lemma D.3. Conditioned on $\mathcal{F}_k$, with probability at least $1 - \delta$, we have

$$\big\|\nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla f_{\mathcal{J}_k}(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1})\big\|_2 \leq 6L\sqrt{\frac{\log(1/\delta)}{B_k^{(g)}}}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2. \tag{D.3}$$

We also have

$$\|\nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla F(\mathbf{x}_k)\|_2 \leq 6M\sqrt{\frac{\log(1/\delta)}{B_k^{(g)}}}. \tag{D.4}$$

Proof of Lemma B.3. First, we have $\mathbf{v}_t - \nabla F(\mathbf{x}_t) = \sum_{k=\lfloor t/S^{(g)}\rfloor \cdot S^{(g)}}^{t} \mathbf{u}_k$, where

$$\begin{aligned} \mathbf{u}_k &= \nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla f_{\mathcal{J}_k}(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1}), && k > \lfloor t/S^{(g)}\rfloor \cdot S^{(g)}, \\ \mathbf{u}_k &= \nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla F(\mathbf{x}_k), && k = \lfloor t/S^{(g)}\rfloor \cdot S^{(g)} \end{aligned}$$

Meanwhile, we have $\mathbb{E}[\mathbf{u}_k|\mathcal{F}_{k-1}] = 0$. Conditioned on $\mathcal{F}_{k-1}$, for $\mathrm{mod}(k, S^{(g)}) \neq 0$, from Lemma D.3, we have that with probability at least $1-\delta$ the following inequality holds :

$$\|\mathbf{u}_k\|_2 \leq 6L\sqrt{\frac{\log(1/\delta)}{B_k^{(g)}}}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 \leq \sqrt{\frac{\epsilon^2}{540 S^{(g)}\log(1/\delta)}}, \tag{D.5}$$

where the second inequality holds due to (4.1). For $\mathrm{mod}(k, S^{(g)}) = 0$, with probability at least $1-\delta$, we have

$$\|\mathbf{u}_k\|_2 \leq 6M\sqrt{\frac{\log(1/\delta)}{B_k^{(g)}}} \leq \frac{\epsilon}{\sqrt{540\log(1/\delta)}}, \tag{D.6}$$

where the second inequality holds due to (4.3). Conditioned on $\mathcal{F}_{\lfloor t/S^{(g)}\rfloor \cdot S^{(g)}}$, by union bound, with probability at least $1 - \delta \cdot (t - \lfloor t/S^{(g)}\rfloor \cdot S^{(g)})$ (D.5) or (D.6) holds for all $\lfloor t/S^{(g)}\rfloor \cdot S^{(g)} \leq k \leq t$. Then for given k, by vector Azuma-Hoeffding inequality in Lemma G.1, conditioned on$\mathcal{F}_k$, with probability at least $1-\delta$ we have

$$\begin{aligned} \|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|_2^2 &= \bigg\|\sum_{k=\lfloor t/S^{(g)}\rfloor \cdot S^{(g)}}^{t} \mathbf{u}_k\bigg\|_2^2 \\ &\leq 9\log(d/\delta)\bigg[(t - \lfloor t/S^{(g)}\rfloor \cdot S^{(g)}) \cdot \frac{\epsilon^2}{540 S^{(g)}\log(d/\delta)} + \frac{\epsilon^2}{540\log(1/\delta)}\bigg] \\ &\leq 9\log(1/\delta) \cdot \frac{\epsilon^2}{270\log(1/\delta)} \\ &\leq \epsilon^2/30. \end{aligned} \tag{D.7}$$

Finally, by union bound, we have that with probability at least $1 - 2\delta\cdot(t - \lfloor t/S^{(g)}\rfloor \cdot S^{(g)})$, for all $\lfloor t/S^{(g)}\rfloor \cdot S^{(g)} \leq k \leq t$, we have (D.7) holds. □

D.4 Proof of Lemma B.4

We need the following lemma:

Lemma D.4. Conditioned on $\mathcal{F}_k$, with probability at least $1-\delta$, we have the following concentration inequality

$$\big\|\nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_k) - \nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_{k-1}) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1})\big\|_2 \leq 6\rho\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2. \tag{D.8}$$

We also have

$$\|\nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_k)\|_2 \leq 6L\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}}. \tag{D.9}$$

Proof of Lemma B.4. First, we have $\mathbf{U}_t - \nabla^2 F(\mathbf{x}_t) = \sum_{k=\lfloor t/S^{(h)}\rfloor \cdot S^{(h)}}^{t} \mathbf{V}_k$, where

$$\begin{aligned} \mathbf{V}_k &= \nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_k) - \nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_{k-1}) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1}), && k > \lfloor t/S^{(h)}\rfloor \cdot S^{(h)}, \\ \mathbf{V}_k &= \nabla f_{\mathcal{I}_k}(\mathbf{x}_k) - \nabla F(\mathbf{x}_k), && k = \lfloor t/S^{(h)}\rfloor \cdot S^{(h)} \end{aligned}$$

Meanwhile, we have $\mathbb{E}[\mathbf{V}_k|\sigma(\mathbf{V}_{k-1}, ..., \mathbf{V}_0)] = 0$. Conditioned on $\mathcal{F}_{k-1}$, for $\mathrm{mod}(k, S^{(h)}) \neq 0$, from Lemma D.4, we have that with probability at least $1-\delta$, the following inequality holds :

$$\|\mathbf{V}_k\|_2 \leq 6\rho\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 \leq \sqrt{\frac{\rho\epsilon}{360 S^{(h)}\log(d/\delta)}}, \tag{D.10}$$

where the second inequality holds due to (4.1). For $\mathrm{mod}(k, S^{(h)}) = 0$, with probability at least $1-\delta$, we have

$$\|\mathbf{V}_k\|_2 \leq 6L\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}} \leq \sqrt{\frac{\rho\epsilon}{360\log(d/\delta)}}, \tag{D.11}$$

where the second inequality holds due to (4.3). Conditioned on $\mathcal{F}_{\lfloor t/S^{(h)}\rfloor\cdot S^{(h)}}$, by union bound, with probability at least $1-\delta\cdot(t-\lfloor t/S^{(h)}\rfloor\cdot S^{(h)})$ (D.10) or (D.11) holds for all $\lfloor t/S^{(h)}\rfloor\cdot S^{(h)} \leq k \leq t$. Then for given k, by Matrix Azuma inequality Lemma G.2, conditioned on$\mathcal{F}_k$, with probability at least $1-\delta$ we have

$$\begin{aligned}
\|\mathbf{U}_k - \nabla^2 F(\mathbf{x}_k)\|_2^2 &= \Bigg\|\sum_{k=\lfloor t/S^{(h)}\rfloor\cdot S^{(h)}}^{t} \mathbf{V}_k\Bigg\|_2^2 \\
&\leq 9\log(d/\delta)\bigg[(t-\lfloor t/S^{(h)}\rfloor\cdot S^{(h)})\cdot\frac{\rho\epsilon}{360S^{(h)}\log(d/\delta)} + \frac{\rho\epsilon}{360\log(d/\delta)}\bigg] \\
&\leq 9\log(d/\delta)\cdot\frac{\rho\epsilon}{180\log(d/\delta)} \\
&\leq \rho\epsilon/20.
\end{aligned} \tag{D.12}$$

Finally, by union bound, we have that with probability at least $1-2\delta\cdot(t-\lfloor t/S^{(h)}\rfloor\cdot S^{(h)})$, for all $\lfloor t/S^{(h)}\rfloor\cdot S^{(h)} \leq k \leq t$, we have (D.12) holds.

□

E Proofs of Technical Lemmas in Appendix C

E.1 Proof of Lemma C.3

We have the following lemma which guarantees the effectiveness of Cubic-Subsolver in Algorithm 3.

Lemma E.1. (Carmon and Duchi, 2016) Let $\mathbf{A}\in\mathbb{R}^{d\times d}$ and $\|\mathbf{A}\|_2\leq\beta$, $\mathbf{b}\in\mathbb{R}^d$, $\tau>0, \zeta>0, \epsilon'\in(0,1), \delta'\in(0,1)$ and $\eta<1/(8\beta+2\tau\zeta)$. We denote that $g(\mathbf{h}) = \mathbf{b}^\top\mathbf{h}+\mathbf{h}^\top\mathbf{A}\mathbf{h}/2+\tau/6\cdot\|\mathbf{h}\|_2^3$ and $\mathbf{s} = \mathrm{argmin}_{\mathbf{h}\in\mathbb{R}^d} g(\mathbf{h})$. Then with probability at least $1-\delta'$, if

$$\|\mathbf{s}\|_2 \geq \zeta \text{ or } \|\mathbf{b}\|_2 \geq \max\{\sqrt{\beta\tau/2}\zeta^{3/2}, \tau\zeta^2/2\}, \tag{E.1}$$

then $\mathbf{x} = \text{Cubic-Subsolver}(\mathbf{A}, \mathbf{b}, \tau, \eta, \zeta, \epsilon', \delta')$ satisfies that $g(\mathbf{x}) \leq -(1-\epsilon')\tau\zeta^3/12$.

Proof of Lemma C.3. We simply set $\mathbf{A} = \mathbf{U}_t$, $\mathbf{b} = \mathbf{v}_t$, $\tau = M_t$, $\eta = (16L)^{-1}$, $\zeta = \sqrt{\epsilon/\rho}$, $\epsilon' = 0.5$ and $\delta' = \delta$. We have $\|\mathbf{U}_t\|_2 \leq L$, then we set $\beta = L$. With the choice of M_t where $M_t = 4\rho$ and the assumption that $\epsilon < 4L^2\rho/M_t^2$, we can check that $\eta < 1/(8\beta+2\tau\zeta)$. We also have that $\mathbf{s} = \mathbf{h}_t^*$ and (E.1) holds. Thus, by Lemma E.1, we have

$$m_t(\mathbf{h}_t) \leq -(1-\epsilon')\tau\zeta^3/12 \leq -M_t\rho^{-3/2}\epsilon^{3/2}/24.$$

By the choice of T' in Cubic-Subsolver, we have

$$T' = \frac{480}{\eta\tau\zeta\epsilon'}\bigg[6\log\Big(1+\sqrt{d}/\delta'\Big) + 32\log\bigg(\frac{12}{\eta\tau\zeta\epsilon'}\bigg)\bigg)\bigg] = \widetilde{O}\bigg(\frac{L}{M_t\sqrt{\epsilon/\rho}}\bigg).$$

□

E.2 Proof of Lemma C.5

We have the following lemma which provides the guarantee for the function value in Cubic-Finalsolver.

Lemma E.2. (Carmon and Duchi, 2016) We denote that $g(\mathbf{h}) = \mathbf{b}^\top\mathbf{h}+\mathbf{h}^\top\mathbf{A}\mathbf{h}/2+\tau/6\cdot\|\mathbf{h}\|_2^3$, $\mathbf{s} = \mathrm{argmin}_{\mathbf{h}\in\mathbb{R}^d} g(\mathbf{h})$, then $g(\mathbf{s}) \geq \|\mathbf{b}\|_2\|\mathbf{s}\|_2/2 - \tau\|\mathbf{s}\|_2^3/6$.

Proof of Lemma C.5. In Cubic-Finalsolver we are focusing on minimizing $m_{T^*-1}(\mathbf{h})$. We have that $\|\mathbf{v}_t\|_2 < \max\{M_t\epsilon/(2\rho), \sqrt{LM_t/2}(\epsilon/\rho)^{3/4}\}$ and $\|\mathbf{h}^*_{T^*-1}\|_2 \leq \sqrt{\epsilon/\rho}$ by Lemma C.3. We can check that $\eta = (16L)^{-1}$ satisfies that $\eta < (4(L+\tau R))^{-1}$, where R is defined in Lemma C.4, when $\epsilon < 4L^2\rho/M_t^2$. From Lemma C.4 we also know that m_{T^*-1} is $(L + 2M_{T^*-1}R)$-smooth, which satisfies that $1/\eta > 2(L + 2M_{T^*-1}R)$. Thus, by standard gradient descent analysis, to get a point Δ where $\|\nabla m_{T^*-1}(\Delta)\|_2 \leq \epsilon$, Cubic-Finalsolver needs to run

$$T'' = O\bigg(\frac{m_{T^*-1}(\Delta_0) - m_{T^*-1}(\mathbf{h}^*_{T^*-1})}{\eta\epsilon^2}\bigg) = O\bigg(L\frac{m_{T^*-1}(\Delta_0) - m_{T^*-1}(\mathbf{h}^*_{T^*-1})}{\epsilon^2}\bigg) \tag{E.2}$$

iterations, where we denote by Δ_0 the starting point of Cubic-Finalsolver. By directly computing, we have $m_{T^*-1}(\Delta_0) \leq 0$. By Lemma E.2, we have

$$-m_{T^*-1}(\mathbf{h}^*_{T^*-1}) \leq M_t\|\mathbf{h}^*_{T^*-1}\|_2^3/6 - \|\mathbf{v}_{T^*-1}\|_2\|\mathbf{h}^*_{T^*-1}\|_2/2 \leq M_t\|\mathbf{h}^*_{T^*-1}\|_2^3/6 = O\big(\rho(\epsilon/\rho)^{3/2}\big) = O(\epsilon^{3/2}/\sqrt{\rho}).$$

Thus, (E.2) can be further bounded as $T'' = O(L/\sqrt{\rho\epsilon})$. □

F Proofs of Additional Lemmas in Appendix D

F.1 Proof of Lemma D.3

Proof of Lemma D.3. We only need to consider the case where $B_k^{(g)} = |\mathcal{J}_k| < n$. For each $i \in \mathcal{J}_k$, let

$$\mathbf{a}_i = \nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1}), \tag{F.1}$$

then we have $\mathbb{E}_i\mathbf{a}_i = 0$, $\mathbf{a}_i$ i.i.d., and

$$\|\mathbf{a}_i\|_2 \leq \|\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1})\|_2 + \|\nabla F(\mathbf{x}_k) - \nabla F(\mathbf{x}_{k-1})\|_2 \leq 2L\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2,$$

where the second inequality holds due to the L-smoothness of f_i and F. Thus by vector Azuma-Hoeffding inequality in Lemma G.1, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\big\|\nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla f_{\mathcal{J}_k}(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1})\big\|_2 \\
&= \frac{1}{B_k^{(g)}}\bigg\|\sum_{i\in\mathcal{J}_k}\Big[\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1}) - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_{k-1})\Big]\bigg\|_2 \\
&\leq 6L\sqrt{\frac{\log(d/\delta)}{B_k^{(g)}}}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2.
\end{aligned} \tag{F.2}$$

For each $i \in \mathcal{J}_k$, let

$$\mathbf{b}_i = \nabla f_i(\mathbf{x}_k) - \nabla F(\mathbf{x}_k),$$

then we have $\mathbb{E}_i\mathbf{b}_i = 0$ and $\|\mathbf{b}_i\|_2 \leq M$. Thus by vector Azuma-Hoeffding inequality in Lemma G.1, we have that with probability at least $1-\delta$,

$$\|\nabla f_{\mathcal{J}_k}(\mathbf{x}_k) - \nabla F(\mathbf{x}_k)\|_2 = \frac{1}{B_k^{(g)}}\bigg\|\sum_{i\in\mathcal{J}_k}\Big[\nabla f_i(\mathbf{x}_k) - \nabla F(\mathbf{x}_k)\Big]\bigg\|_2 \leq 6M\sqrt{\frac{\log(d/\delta)}{B_k^{(g)}}}. \tag{F.3}$$

□

F.2 Proof of Lemma D.4

Proof of Lemma D.4. We only need to consider the case where $B_k^{(h)} = |\mathcal{I}_k| < n$. For each $i \in \mathcal{I}_k$, let

$$\mathbf{A}_i = \nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\mathbf{x}_{k-1}) - \nabla^2 F(\mathbf{x}_k) + \nabla^2 F(\mathbf{x}_{k-1}),$$

then we have $\mathbb{E}_i\mathbf{A}_i = 0, \mathbf{A}_i^\top = \mathbf{A}_i$, $\mathbf{A}_i$ i.i.d. and

$$\|\mathbf{A}_i\|_2 \leq \big\|\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\mathbf{x}_{k-1})\big\|_2 + \big\|\nabla^2 F(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_{k-1})\big\|_2 \leq 2\rho\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2,$$

where the second inequality holds due to ρ-Hessian Lipschitz continuous of f_i and F. Then by Matrix Azuma inequality Lemma G.2, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\big\|\nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_k)-\nabla^2 f_{\mathcal{I}_k}(\mathbf{x}_{k-1})-\nabla^2 F(\mathbf{x}_k)+\nabla^2 F(\mathbf{x}_{k-1})\big\|_2\\
&=\frac{1}{B_k^{(h)}}\bigg\|\sum_{i\in\mathcal{I}_k}\Big[\nabla^2 f_i(\mathbf{x}_k)-\nabla^2 f_i(\mathbf{x}_{k-1})-\nabla^2 F(\mathbf{x}_k)+\nabla^2 F(\mathbf{x}_{k-1})\Big]\bigg\|_2\\
&\le 6\rho\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}}\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2.
\end{aligned}$$

For each $i\in\mathcal{I}_k$, let

$$\mathbf{B}_i=\nabla^2 f_i(\mathbf{x}_k)-\nabla^2 F(\mathbf{x}_k),$$

then we have $\mathbb{E}_i\mathbf{B}_i=0$, $\mathbf{B}_i^\top=\mathbf{B}_i$, and $\|\mathbf{B}_i\|_2\le 2L$. Then by Matrix Azuma inequality in Lemma G.2, we have that with probability at least $1-\delta$,

$$\|\nabla^2 f_{\mathcal{J}_k}(\mathbf{x}_k)-\nabla^2 F(\mathbf{x}_k)\|_2=\frac{1}{B_k^{(h)}}\bigg\|\sum_{i\in\mathcal{I}_k}\Big[\nabla^2 f_i(\mathbf{x}_k)-\nabla^2 F(\mathbf{x}_k)\Big]\bigg\|_2\le 6L\sqrt{\frac{\log(d/\delta)}{B_k^{(h)}}},$$

which completes the proof. □

G Auxiliary Lemmas

We have the following vector Azuma-Hoeffding inequality:

Lemma G.1. (Pinelis, 1994) Consider $\{\mathbf{v}_k\}$ be a vector-valued martingale difference, where $\mathbb{E}[\mathbf{v}_k|\sigma(\mathbf{v}_1,...,\mathbf{v}_{k-1})]=0$ and $\|\mathbf{v}_k\|_2\le A_k$, then we have that with probability at least $1-\delta$,

$$\bigg\|\sum_k\mathbf{v}_k\bigg\|_2\le 3\sqrt{\log(1/\delta)\sum_k A_k^2} \tag{G.1}$$

We have the following Matrix Azuma inequality :

Lemma G.2. (Tropp, 2012) Consider a finite adapted sequence $\{\mathbf{X}_k\}$ of self-adjoint matrices in dimension d, and a fixed sequence $\{\mathbf{A}_k\}$ of self-adjoint matrices that satisfy

$$\mathbb{E}[\mathbf{X}_k|\sigma(\mathbf{X}_{k-1},...,\mathbf{X}_1)]=\mathbf{0}\text{ and }\mathbf{X}_k^2\preceq\mathbf{A}_k^2\text{ almost surely.}$$

Then we have that with probability at least $1-\delta$,

$$\bigg\|\sum_k\mathbf{X}_k\bigg\|_2\le 3\sqrt{\log(d/\delta)\sum_k\|\mathbf{A}_k\|_2^2}. \tag{G.2}$$

H Additional Algorithms and Functions

Due to space limit, we include the approximate solvers (Carmon and Duchi, 2016) for the cubic subproblem in this section for the purpose of self-containedness.

Algorithm 3 Cubic-Subsolver($\mathbf{A}[\cdot], \mathbf{b}, \tau, \eta, \zeta, \epsilon', \delta'$)

```
$\mathbf{x} = \text{CauchyPoint}(\mathbf{A}[\cdot], \mathbf{b}, \tau)$
if CubicFunction($\mathbf{A}[\cdot], \mathbf{b}, \tau, \mathbf{x}$) $\le -(1-\epsilon')\tau\zeta^3/12$ then
    return x
end if
Set
    $$T' = \frac{480}{\eta\tau\zeta\epsilon'}\bigg[6\log\Big(1+\sqrt{d}/\delta'\Big) + 32\log\bigg(\frac{12}{\eta\tau\zeta\epsilon'}\bigg)\bigg)\bigg]$$
Draw $\mathbf{q}$ uniformly from the unit sphere, set $\widetilde{\mathbf{b}} = \mathbf{b} + \sigma\mathbf{q}$ where $\sigma = \tau^2\zeta^3\epsilon'/(\beta+\tau\zeta)/576$
$\mathbf{x} = \text{CauchyPoint}(\mathbf{A}[\cdot], \mathbf{b}, \tau)$
for $t = 1, \ldots, T-1$ do
    $\mathbf{x} \leftarrow \mathbf{x} - \eta \cdot \text{CubicGradient}(\mathbf{A}[\cdot], \widetilde{\mathbf{b}}, \tau, \mathbf{x})$
    if CubicFunction($\mathbf{A}[\cdot], \widetilde{\mathbf{b}}, \tau, \mathbf{x}$) $\le -(1-\epsilon')\tau\zeta^3/12$ then
        return x
    end if
end for
return x
```

Algorithm 4 Cubic-Finalsolver($\mathbf{A}[\cdot], \mathbf{b}, \tau, \eta, \epsilon_g$)

```
$\Delta \leftarrow$CauchyPoint($\mathbf{A}[\cdot], \mathbf{b}, \tau$)
while $\|\text{Gradient}(\mathbf{A}[\cdot], \mathbf{b}, \tau, \Delta)\|_2 > \epsilon_g$ do
    $\Delta \leftarrow \Delta - \eta \cdot \text{Gradient}(\mathbf{A}[\cdot], \mathbf{b}, \tau, \Delta)$
end while
return $\Delta$
```

```
Function: CauchyPoint($\mathbf{A}[\cdot], \mathbf{b}, \tau$)
return $-R_c\mathbf{b}/\|\mathbf{b}\|_2$, where
    $$R_c = \frac{-\mathbf{b}^\top\mathbf{A}[\mathbf{b}]}{\tau\|\mathbf{b}\|_2^2} + \sqrt{\bigg(\frac{-\mathbf{b}^\top\mathbf{A}[\mathbf{b}]}{\tau\|\mathbf{b}\|_2^2}\bigg)^2 + \frac{2\|\mathbf{b}\|_2}{\tau}}$$

Function: CubicFunction($\mathbf{A}[\cdot], \mathbf{b}, \tau, \mathbf{x}$)
return $\mathbf{b}^\top\mathbf{x} + \mathbf{x}^\top\mathbf{A}[\mathbf{x}]/2 + \tau\|\mathbf{x}\|_2^3/6$

Function: CubicGradient($\mathbf{A}[\cdot], \mathbf{b}, \tau, \mathbf{x}$)
return $\mathbf{b}^\top + \mathbf{A}[\mathbf{x}] + \tau\|\mathbf{x}\|_2\mathbf{x}/2$
```