

Change this to the title of your paper *

FRANK E. CURTIS[†], DANIEL P. ROBINSON[‡], AND LINGJUN GUO[§]

Abstract. Add abstract here.

Key words. nonlinear optimization, nonconvex optimization, worst-case iteration complexity, worst-case evaluation complexity, regularization methods, trust region methods

AMS subject classifications. 49M37, 65K05, 65K10, 65Y20, 68Q25, 90C30, 90C60

1. Introduction. Equality-constrained optimization problems arise...

[Lingjun](#): Add a citation to the paper for the unconstrained setting. The unconstrained progressive sampling paper is [1].

1.1. Contributions. Our contributions relate ...

1.2. Notation. We use \mathbb{R} to denote the set of real numbers, $\mathbb{R}_{\geq r}$ (resp., $\mathbb{R}_{>r}$) to denote the set of real numbers greater than or equal to (resp., greater than) $r \in \mathbb{R}$, \mathbb{R}^n to denote the set of n -dimensional real vectors, and $\mathbb{R}^{m \times n}$ to denote the set of m -by- n -dimensional real matrices. We denote the set of nonnegative integers as $\mathbb{N} := \{0, 1, 2, \dots\}$, and, for any integer $N \geq 1$, we use $[N]$ to denote the set $\{1, \dots, N\}$.

For any finite set \mathcal{S} , we use $|\mathcal{S}|$ to denote its cardinality. We consider all vector norms to be Euclidean, i.e., we let $\|\cdot\| := \|\cdot\|_2$, unless otherwise specified. Similarly, we use $\|\cdot\|$ to denote the spectral norm of any matrix input.

For any matrix $A \in \mathbb{R}^{m \times n}$, we use $\sigma_i(A)$ to denote its i th largest singular value. Given any such A , we use $\text{Null}(A)$ to denote its null space, i.e., $\{d \in \mathbb{R}^n : Ad = 0\}$. Assuming $B \in \mathbb{R}^{n \times m}$ has full column rank, we use B^\dagger to denote its pseudoinverse, i.e., $B^\dagger := (B^T B)^{-1} B^T$. For any subspace $\mathcal{X} \subseteq \mathbb{R}^n$ and point $x \in \mathbb{R}^n$, we denote the projection of x onto \mathcal{X} as $\text{Proj}_{\mathcal{X}}(x) := \arg \min_{\bar{x} \in \mathcal{X}} \|\bar{x} - x\|$. Given $B \in \mathbb{R}^{n \times m}$ with full column rank, we use $\mathcal{R}(B) := BB^\dagger$ and $\mathcal{N}(B) = I - \mathcal{R}(B)$ to denote projection matrices onto the span of the columns of B and the null space of B , respectively.

1.3. Organization. In §3, ...

2. Algorithm. Our proposed algorithm is designed to solve a sample average approximation (SAA) of the continuous nonlinear-equality-constrained problem

$$(2.1) \quad \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } \bar{c}(x) = 0,$$

where the objective and constraint functions, i.e., $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\bar{c} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, respectively, are continuously differentiable, $m \leq n$, and the constraint function c is defined by an expectation. Formally, with respect to a random variable ω defined by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the expectation function \mathbb{E} defined by \mathbb{P} , and $\bar{C} : \mathbb{R}^n \times \Omega \rightarrow$

*This material is based upon work supported by the U.S. Department of Energy, Office of Science, Applied Mathematics, Early Career Research Program under Award Number DE-SC0010615 and by the U.S. National Science Foundation, Division of Mathematical Sciences, Computational Mathematics Program under Award Number DMS-1016291.

[†]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA;
E-mail: frank.e.curtis@lehigh.edu

[‡]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA;
E-mail: daniel.p.robinson@lehigh.edu

[§]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA;
E-mail: lig423@lehigh.edu

34 \mathbb{R}^m , the constraint function \bar{c} is defined by $\bar{c}(x) = \mathbb{E}[\bar{C}(x, \omega)]$ for all $x \in \mathbb{R}^n$. The
 35 SAA of problem (2.1) that our algorithm is designed to solve is defined with respect
 36 to a sample of $N \in \mathbb{N}$ realizations of the random variable ω , say, $\{\omega_i\}_{i \in [N]}$. Defining
 37 the SAA constraint function $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for all $x \in \mathbb{R}^n$ by

$$38 \quad c(x) = \frac{1}{N} \sum_{i=1}^N c_i(x), \quad \text{where } c_i(x) \equiv \bar{C}(x, \omega_i) \quad \text{for all } i \in [N],$$

39 the problem that our algorithm is designed to solve is that given by

$$40 \quad (2.2) \quad \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } c(x) = 0.$$

41 Under mild assumptions about c and an assumption that N is sufficiently large, a
 42 point that is approximately stationary for problem (2.2) can be shown to be approxi-
 43 mately stationary for problem (2.1), at least with high probability. We leave a formal
 44 statement and proof of this fact until the end of our analysis. Until that time, we
 45 focus on our proposed algorithm and our analysis of it for solving problem (2.2).

46 The main idea of our proposed algorithm for solving problem (2.2) is to generate
 47 a sequence of iterates, each of which is a stationary point (at least approximately)
 48 with respect to a subsampled problem involving only a subset $\mathcal{S} \subseteq [N]$ of constraint
 49 function terms. For any such \mathcal{S} , an approximation of problem (2.2) is given by

$$50 \quad (2.3) \quad \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } c_{\mathcal{S}}(x) = 0, \quad \text{where } c_{\mathcal{S}}(x) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} c_i(x).$$

51 The primary benefit of considering (2.3) for $\mathcal{S} \subseteq [N]$, rather than (2.2) directly, is
 52 that any evaluation of a constraint or constraint Jacobian value requires computing
 53 a sum of $|\mathcal{S}| \leq N$ terms, as opposed to N terms. Also, under assumptions about
 54 the constraint functions that are reasonable for many real-world problems of interest,
 55 we show in this paper that, by starting with an approximate stationary point for
 56 problem (2.3) and aiming to solve a subsequent instance of (2.3) with respect to a
 57 sample set $\bar{\mathcal{S}} \supseteq \mathcal{S}$, our proposed algorithm can obtain an approximate stationary
 58 point for the subsequent instance with lower sample complexity than if the problem
 59 with the larger sample set were solved directly. Overall, we show that—at least once
 60 the sample sets become sufficiently large relative to N —a sufficiently approximate
 61 stationary point of problem (2.2) can be obtained more efficiently through progressive
 62 sampling than by tackling the problem directly.

63 For use in our proposed algorithm and our analysis of it, let us introduce sta-
 64 tionarity conditions for problem (2.3), which also represent stationarity conditions for
 65 problem (2.2) in the particular case when $\mathcal{S} = [N]$. The Lagrangian of problem (2.3)
 66 is $L_{\mathcal{S}} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ defined for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ by

$$67 \quad L_{\mathcal{S}}(x, y) = f(x) + c_{\mathcal{S}}(x)^T y = f(x) + \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} c_i(x)^T y,$$

68 where $y \in \mathbb{R}^m$ is referred to as a vector of Lagrange multipliers or dual variables.
 69 Second-order necessary conditions for optimality for (2.3) can then be stated as

$$70 \quad (2.4a) \quad \nabla_x L_{\mathcal{S}}(x, y) = 0$$

$$71 \quad (2.4b) \quad \text{and } d^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y) d \geq 0 \quad \text{for all } d \in \text{Null}(\nabla c_{\mathcal{S}}(x)^T).$$

We refer to any point (x, y) satisfying (2.4a) as a first-order stationary point with respect to problem (2.3), and we refer to any point satisfying both (2.4a) and (2.4b) (i.e., satisfying (2.4)) as a second-order stationary point with respect to problem (2.3). In addition, consistent with the literature on worst-case complexity bounds for non-convex smooth optimization, we say that a point (x, y) is (ϵ, ε) -stationary with respect to problem (2.3) for some $(\epsilon, \varepsilon) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ if and only if

$$(2.5a) \quad \|\nabla_x L_{\mathcal{S}}(x, y)\| \leq \epsilon$$

$$(2.5b) \quad \text{and } d^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y) d \geq -\varepsilon \text{ for all } d \in \text{Null}(\nabla c_{\mathcal{S}}(x)^T).$$

Generally speaking, an algorithm for solving (2.3) can be a *primal* method that might only generate a sequence of primal iterates $\{x_k\}$, or it can be a *primal-dual* method that generates a sequence of primal and dual iterate pairs $\{(x_k, y_k)\}$. For an application of our proposed algorithm, either type of method can be employed, but for certain results in our analysis we refer to properties of *least-square multipliers* corresponding to a given primal point $x \in \mathbb{R}^n$. Assuming that the Jacobian of $c_{\mathcal{S}}$ at x , namely, $\nabla c_{\mathcal{S}}(x)^T$, has full row rank, the least-squares multipliers with respect to x are given by $y(x) \in \mathbb{R}^m$ that minimizes $\|\nabla_x L(x, \cdot)\|^2$, which is given by

$$(2.6) \quad y(x) = -(\nabla c_{\mathcal{S}}(x)^T \nabla c_{\mathcal{S}}(x))^{-1} \nabla c_{\mathcal{S}}(x)^T \nabla f(x) = -\nabla c_{\mathcal{S}}(x)^{\dagger} \nabla f(x).$$

Our proposed method is stated as Algorithm 2.1 below.

Algorithm 2.1 Progressive Constraint-Sampling Method (PCSM) for (2.2)

Require: Initial sample set size $p_1 \in [N]$, initial point $x_0 \in \mathbb{R}^n$, maximum outer iteration index $K = \lceil \log_2 \frac{N}{p_1} \rceil$, and subproblem tolerances $\{(\epsilon_k, \varepsilon_k)\}_{k=1}^K \subset \mathbb{R}_{>0}$

- 1: set $\mathcal{S}_0 \leftarrow \emptyset$
- 2: **for** $k \in [K]$ **do**
- 3: choose $\mathcal{S}_k \supseteq \mathcal{S}_{k-1}$ such that $|\mathcal{S}_k| = p_k$
- 4: using x_{k-1} as a starting point, employ an algorithm to solve (2.3), terminating once a primal iterate x_k has been obtained such that $(x_k, y(x_k))$ (see (2.6)) is $(\epsilon_k, \varepsilon_k)$ -stationary with respect to problem (2.3) for $\mathcal{S} = \mathcal{S}_k$
- 5: set $p_{k+1} \leftarrow \min\{2p_k, N\}$
- 6: **end for**
- 7: **return** $(x_K, y(x_K))$, which is $(\epsilon_K, \varepsilon_K)$ -stationary with respect to (2.2)

92 3. Analysis. FEC: Moved assumption from earlier....

93 ASSUMPTION 3.1. For all $N \in \mathbb{N}$, there exists a sample size $p_N \in [N]$, such that
 94 for all $(x, \mathcal{S}) \subseteq \mathbb{R}^n \times [N]$ with $|\mathcal{S}| \geq p_N$, the Jacobian $\nabla c_{\mathcal{S}}(x)^T$ is nondegenerate, i.e.
 95 $\text{rank}(\nabla c_{\mathcal{S}}(x)^T) = m$.

96 ASSUMPTION 3.2. There exist constants $(\sigma_c^{\max}, \sigma_f^{\max}, \sigma_c^{\min}, \lambda_c^{\max}, \lambda_f^{\max}) \in \mathbb{R}_{>0}^5$,
 97 such that for all $(j, x) \in [m] \times \mathbb{R}^n$, the following hold

- 98 (1). We have $\|\nabla c(x)\|_2 \leq \sigma_c^{\max}$ and $\|\nabla f(x)\|_2 \leq \sigma_f^{\max}$. Moreover, the smallest
 99 singular value of $\nabla c(x)$, i.e. $\sigma_m(\nabla c(x))$ satisfies $\sigma_m(\nabla c(x)) \geq \sigma_c^{\min}$.
- 100 (2). We have $\|\nabla^2 f(x)\|_2 \leq \lambda_f^{\max}$. In addition, we have $\|\nabla^2 c^j(x)\|_2 \leq \lambda_c^{\max}$ where
 101 the function c^j is the j th element of c .

102 In addition to the above assumptions for the average constraint function c , we also
 103 make the following assumptions regarding individual sample function c_i .

ASSUMPTION 3.3. For all $x \in \mathbb{R}^n$, the Jacobian $\nabla c(x)^T$ has full column rank. In addition, there exist positive constants $(\theta_J, \nu_J, \mu_H) \in \mathbb{R}_{>0}^3$, such that the following hold

(1). For any $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \|\nabla c_i(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T\|_2^2 &\leq \theta_J \|\nabla c(x)^T\|_2^2, \text{ and} \\ \frac{1}{N} \sum_{i=1}^N \|\nabla c_i(x)^T \mathcal{N}(\nabla c(x))\|_2^2 &\leq \nu_J \|\nabla c(x)^T\|_2^2. \end{aligned}$$

(2). For all $(j, x) \in [m] \times \mathbb{R}^n$, we have

$$\frac{1}{N} \sum_{i=1}^N \|\nabla^2 c_i^j(x) - \nabla^2 c^j(x)\|_2^2 \leq \mu_H \|\nabla^2 c^j(x)\|_2^2.$$

We make the following two definitions.

DEFINITION 3.1. For any $(\alpha, \beta) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, problem $\{\min_x f(x), \text{s.t. } c(x) = 0\}$ is (α, β) -morse, if and only if, for any $x \in \mathbb{R}^n$ there exists a $y \in \mathbb{R}^m$, such that when (x, y) satisfies $\|\nabla_x L(x, y)\|_2 \leq \alpha$, we have $|d^T \nabla_{xx}^2 L(x, y) d| \geq \beta \|d\|_2^2$ for all $d \in \text{Null}(\nabla c(x)^T)$.

DEFINITION 3.2. For any full column rank matrices $(A, B) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m}$ where $n \geq m$, the A and B are acute perturbations to each other, if and only if

$$\text{rank}(AA^\dagger BA^\dagger A) = m.$$

LEMMA 3.3. Under Assumption 3.2 where constant σ_c^{\min} exist. In addition, under Assumption 3.3 where constants (θ_J, ν_J, μ_H) exist. Then, the following hold

(1). A sample average result holds, that is, for any $(j, x, \mathcal{S}) \in [m] \times \mathbb{R}^n \times [N]$, we have

$$\begin{aligned} \|\nabla c_{\mathcal{S}}(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T\|_2^2 &\leq N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right) \theta_J \|\nabla c(x)^T\|_2^2, \\ \|\nabla c_{\mathcal{S}}(x)^T \mathcal{N}(\nabla c(x))\|_2^2 &\leq N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right) \nu_J \|\nabla c(x)\|_2^2, \\ \|\nabla^2 c_{\mathcal{S}}^j(x) - \nabla^2 c^j(x)\|_2^2 &\leq N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right) \mu_H \|\nabla^2 c^j(x)\|_2^2. \end{aligned}$$

(2). The ∇c^\dagger , $y_{[N]}$ and $\nabla_{xx}^2 L_{[N]}$ are bounded, that is, for any $x \in \mathbb{R}^n$ we have

$$\begin{aligned} \|\nabla c(x)^\dagger\|_2 &\leq \frac{1}{\sigma_c^{\min}} \text{ and } \|y_{[N]}(x)\|_2 \leq \frac{\sigma_f^{\max}}{\sigma_c^{\min}}, \text{ moreover} \\ \|\nabla_{xx}^2 L_{[N]}(x, y_{[N]})\|_2 &\leq \lambda_f^{\max} + \frac{\sqrt{m} \sigma_f^{\max} \lambda_c^{\max}}{\sigma_c^{\min}}. \end{aligned}$$

Proof. For the first item, we only show the first inequality in (3.1), and the other

127 two inequalities follow a similar argument. Notice that

$$\begin{aligned}
 \nabla c_{\mathcal{S}}(x) &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla c_i(x) = \frac{1}{|\mathcal{S}|} \sum_{i \in [N]} \nabla c_i(x) - \frac{1}{|\mathcal{S}|} \sum_{i \in [N] \setminus \mathcal{S}} \nabla c_i(x) \\
 &= \frac{N}{|\mathcal{S}|} \nabla c(x) - \frac{1}{|\mathcal{S}|} \sum_{i \in [N] \setminus \mathcal{S}} \nabla c_i(x),
 \end{aligned}
 \tag{3.3}$$

129 we have

$$\begin{aligned}
 &\|\nabla c_{\mathcal{S}}(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T\|_2^2 \\
 &= \left\| \frac{N}{|\mathcal{S}|} \nabla c(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T - \frac{1}{|\mathcal{S}|} \sum_{i \in [N] \setminus \mathcal{S}} \nabla c_i(x)^T \mathcal{R}(\nabla c(x)) \right\|_2^2 \\
 &= \underbrace{\left\| \frac{N-|\mathcal{S}|}{|\mathcal{S}|} \nabla c(x)^T - \frac{1}{|\mathcal{S}|} \sum_{i \in [N] \setminus \mathcal{S}} \nabla c_i(x)^T \mathcal{R}(\nabla c(x)) \right\|_2^2}_{(i)}.
 \end{aligned}$$

131 Here, the second line substitutes (3.3) into the equation. For the third line, by the
 132 definition of \mathcal{R} , we have $\nabla c(x)^T \mathcal{R}(\nabla c(x)) = \nabla c(x)^T$, and substitute it to the first
 133 term of the second line gives the result. Further, for (i), we have

$$\begin{aligned}
 (i) &= \frac{1}{|\mathcal{S}|^2} \left\| \sum_{i \in [N] \setminus \mathcal{S}} \{(\nabla c(x)^T - \nabla c_i(x)^T \mathcal{R}(\nabla c(x))) \times I_n\} \right\|_2^2 \\
 &\leq \frac{1}{|\mathcal{S}|^2} \sum_{i \in [N] \setminus \mathcal{S}} \|\nabla c(x)^T - \nabla c_i(x)^T \mathcal{R}(\nabla c(x))\|_2^2 \sum_{i \in [N] \setminus \mathcal{S}} \|I_n\|_2^2 \\
 &= \left(\frac{N-|\mathcal{S}|}{|\mathcal{S}|^2} \right) \sum_{i \in [N] \setminus \mathcal{S}} \|\nabla c(x)^T - \nabla c_i(x)^T \mathcal{R}(\nabla c(x))\|_2^2 \\
 &\leq \left(\frac{N-|\mathcal{S}|}{|\mathcal{S}|^2} \right) \sum_{i \in [N]} \|\nabla c(x)^T - \nabla c_i(x)^T \mathcal{R}(\nabla c(x))\|_2^2 \\
 &\leq \left(\frac{N-|\mathcal{S}|}{|\mathcal{S}|^2} \right) N \theta_J \|\nabla c(x)^T\|_2^2.
 \end{aligned}$$

135 Here, the first line puts the denominator outside the norm and uses a fact that
 136 $(N-|\mathcal{S}|) \nabla c(x)^T = \sum_{i \in [N] \setminus \mathcal{S}} \nabla c_i(x)^T$. The second line uses the Cauchy-Schwarz in-
 137 equality. The third line uses that $\|I_n\|_2 = 1$. The second to last line adds extra $|\mathcal{S}|$
 138 nonnegative terms, and the last line uses the first item of Assumption 3.3.

139 For the second item, see [red](#) for a proof for the bound on $\|\nabla c(x)^\dagger\|_2$.

140 For the bound for $\|y_{[N]}(x)\|_2$, by Assumption 3.2, first item of Lemma 3.3 and
 141 sub-multiplicity for matrix-vector product, we have

$$\|y_{[N]}(x)\|_2 = \|\nabla c(x)^\dagger \nabla f(x)\|_2 \leq \|\nabla c(x)^\dagger\|_2 \|\nabla f(x)\|_2 \leq \frac{\sigma_f^{\max}}{\sigma_c^{\min}}.$$

144 For the last inequality, first, for any $(j, \mathcal{S}) \subseteq [m] \times [N]$ we have

$$\begin{aligned}
\|\nabla^2 c_{\mathcal{S}}^j(x)\|_2 &\leq \|\nabla^2 c^j(x)\|_2 + \|\nabla^2 c^j(x) - \nabla^2 c_{\mathcal{S}}^j(x)\|_2 \\
&\leq \|\nabla^2 c^j(x)\|_2 + \sqrt{\mu_H N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right)} \|\nabla^2 c^j(x)\|_2 \\
&\leq \left(1 + \sqrt{\mu_H N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right)} \right) \lambda_c^{\max}.
\end{aligned}
\tag{3.4}$$

146 Here, the first line adds, subtracts a term, and uses the triangle inequality. The second
147 line uses bound on Hessian in this Lemma. The last inequality uses Assumption 3.2.

148 Second, note that for any vector $y \in \mathbb{R}^m$, we have $\|y\|_1 \leq \sqrt{m}\|y\|_2$. Combining
149 this result with Assumption 3.2, we have

$$\begin{aligned}
\|\nabla_{xx}^2 L_{[N]}(x, y_{[N]})\|_2 &= \|\nabla^2 f(x) + \sum_{j=1}^m y_{[N]}^j \nabla^2 c^j(x)\|_2 \\
&\leq \|\nabla^2 f(x)\|_2 + \left\| \sum_{j=1}^m y_{[N]}^j \nabla^2 c^j(x) \right\|_2 \\
&\leq \lambda_f^{\max} + \max_j \{\|\nabla^2 c^j(x)\|_2\} \|y_{[N]}\|_1 \\
&\leq \lambda_f^{\max} + \sqrt{m} \max_j \{\|\nabla^2 c^j(x)\|_2\} \|y_{[N]}\|_2 \\
&\leq \lambda_f^{\max} + \frac{\sqrt{m} \sigma_f^{\max} \lambda_c^{\max}}{\sigma_c^{\min}}.
\end{aligned}$$

151 Here, the second line uses the triangle inequality. The third line uses multiplicity and
152 $\|\nabla^2 c^j(x)\|_2 \leq \max_j \{\|\nabla^2 c^j(x)\|_2\}$. The rest lines use Assumption 3.2 and the norm
153 relationship. \square

154 With Definition 3.2 and Lemma 3.3, we have the following condition on $(\mathcal{S}, \theta_J, \nu_J)$
155 to ensure the Jacobian $\nabla c(x)^T$ and $\nabla c_{\mathcal{S}}(x)^T$ are acute perturbations to each other.

156 LEMMA 3.4. Under Assumption 3.2 where constants $(\sigma_c^{\min}, \sigma_c^{\max})$ exist. In addi-
157 tion, under Assumption 3.3 where constants (θ_J, ν_J) exist. Then, if $\mathcal{S} \subseteq [N]$ satisfies

$$|\mathcal{S}| > \frac{2}{1 + \sqrt{1 + \frac{2(\sigma_c^{\min})^2}{(\theta_J + \nu_J)(\sigma_c^{\max})^2}}} N,$$

159 the following hold

- 160 (1). For any $x \in \mathbb{R}^n$ the Jacobian $\nabla c_{\mathcal{S}}(x)^T$ is nondegenerate and the associated
161 least square estimator $y_{\mathcal{S}}$ in (2.6) is well-defined.
- 162 (2). For any $x \in \mathbb{R}^n$, the gradient $\nabla c(x)$ and $\nabla c_{\mathcal{S}}(x)$ are acute perturbations to
163 each other.

Proof. First, we examine the difference between $\nabla c_{\mathcal{S}}(x)^T$ and $\nabla c(x)^T$. We have

$$\begin{aligned}
& \|\nabla c_{\mathcal{S}}(x)^T - \nabla c(x)^T\|_2^2 \\
&= \|\nabla c_{\mathcal{S}}(x)^T (\mathcal{R}(\nabla c(x)) + \mathcal{N}(\nabla c(x))) - \nabla c(x)^T\|_2^2 \\
&= \|\nabla c_{\mathcal{S}}(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T + \nabla c_{\mathcal{S}}(x)^T \mathcal{N}(\nabla c(x))\|_2^2 \\
&\leq 2 \|\nabla c_{\mathcal{S}}(x)^T \mathcal{R}(\nabla c(x)) - \nabla c(x)^T\|_2^2 + 2 \|\nabla c_{\mathcal{S}}(x)^T \mathcal{N}(\nabla c(x))\|_2^2 \\
&\leq 2N \left(\frac{N - |\mathcal{S}|}{|\mathcal{S}|^2} \right) (\theta_J + \nu_J) \|\nabla c(x)^T\|_2^2.
\end{aligned}
\tag{3.5}$$

Here, the second line uses $I_n = \mathcal{R}(\nabla c(x)) + \mathcal{N}(\nabla c(x))$. The third line rearranges terms. The second to last line uses the Cauchy-Schwarz inequality, and the last line uses Lemma 3.3.

Further, [3, Theorem 1] gives us a bound on the difference of the smallest singular values, i.e. $|\sigma_m(\nabla c_{\mathcal{S}}(x)^T) - \sigma_m(\nabla c(x)^T)| \leq \|\nabla c_{\mathcal{S}}(x) - \nabla c(x)\|_2$. Combining it with (3.5) we have

$$|\sigma_m(\nabla c_{\mathcal{S}}(x)^T) - \sigma_m(\nabla c(x)^T)| \leq \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \|\nabla c(x)^T\|_2.
\tag{3.6}$$

By the choice of \mathcal{S} , we have $\sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} < \frac{\sigma_c^{\min}}{\sigma_c^{\max}}$, which gives a bound for the smallest singular value of $\nabla c_{\mathcal{S}}(x)^T$,

$$\begin{aligned}
\sigma_m(\nabla c_{\mathcal{S}}(x)^T) &= \sigma_m(\nabla c(x)^T) + \sigma_m(\nabla c_{\mathcal{S}}(x)^T) - \sigma_m(\nabla c(x)^T) \\
&\geq \sigma_m(\nabla c(x)^T) - |\sigma_m(\nabla c_{\mathcal{S}}(x)^T) - \sigma_m(\nabla c(x)^T)| \\
&\geq \sigma_m(\nabla c(x)^T) - \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \|\nabla c(x)\|_2 \\
&> \sigma_c^{\min} - \sigma_c^{\min} = 0.
\end{aligned}$$

Here, the first line adds and subtracts a term. The third line plugs in (3.6). The above result indicates that the smallest singular value of $\nabla c_{\mathcal{S}}(x)^T$ is positive, and we can conclude that $\nabla c_{\mathcal{S}}(x)^T$ is of full column rank and the dual variable $y(x)$ in (2.6) is well defined.

For the second item, by Assumption 3.2, for any $x \in \mathbb{R}^n$, we have the Jacobian $\nabla c(x)^T$ is of full row rank. Further, we have

$$\begin{aligned}
& \nabla c(x) \nabla c(x)^\dagger \nabla c_{\mathcal{S}}(x) \nabla c(x)^\dagger \nabla c(x) \\
&= \nabla c(x) \nabla c(x)^\dagger \nabla c_{\mathcal{S}}(x) \\
&= \nabla c(x) \nabla c(x)^\dagger (\nabla c(x) + \nabla c_{\mathcal{S}}(x) - \nabla c(x)) \\
&= \nabla c(x) \left(I_m + \underbrace{\nabla c(x)^\dagger (\nabla c_{\mathcal{S}}(x) - \nabla c(x))}_{(ii)} \right).
\end{aligned}$$

Here, the second line uses the definition of pseudo-inverse that $\nabla c(x)^\dagger \nabla c(x) = I_m$. The second to last line adds and subtracts a term, and the last line combines the product of the last two terms from the previous equality.

Combining the sub-multiplicity of the matrix product, the first item of Lemma 3.3, inequality (3.5) and the choice of \mathcal{S} , we have:

$$\begin{aligned} \|(ii)\|_2 &= \|\nabla c(x)^\dagger (\nabla c_{\mathcal{S}}(x) - \nabla c(x))\|_2 \\ &\leq \|\nabla c(x)^\dagger\|_2 \|\nabla c_{\mathcal{S}}(x) - \nabla c(x)\|_2 \\ &\leq \frac{1}{\sigma_c^{\min}} \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \|\nabla c(x)\|_2 \\ &< \frac{1}{\sigma_c^{\min}} \sigma_c^{\min} = 1. \end{aligned}$$

Again, by [3, Theorem 1], we have

$$\sigma_m(I_m) - \sigma_m(I_m + (ii)) \leq |\sigma_m(I_m) - \sigma_m(I_m + (ii))| \leq \|(ii)\|_2 < 1,$$

and the most left and right terms of the above inequality give us

$$\sigma_m(I_m + (ii)) > \sigma_m(I_m) - 1 = 1 - 1 = 0,$$

which is positive. Hence, the matrix $I_m + (ii)$ is of full rank. Combining with the fact that $\nabla c_{[N]}(x)^T$ has full column rank, we have that $\nabla c(x)^T(I_m + (ii))$ has full column rank, that gives us

$$\text{rank}(\nabla c(x)\nabla c(x)^\dagger \nabla c_{\mathcal{S}}(x)\nabla c(x)^\dagger \nabla c(x)) = m.$$

By Definition 3.2, the $\nabla c(x)$ and $\nabla c_{\mathcal{S}}(x)$ are acute perturbations to each other. \square

Now, we can present the first type of bounds.

LEMMA 3.5. *Under Assumption 3.2 where constants $(\sigma_f^{\max}, \sigma_c^{\min}, \sigma_c^{\max})$ exist. In addition, under Assumption 3.3 where constants (θ_J, ν_J) exist. Then, for any $x \in \mathbb{R}^n$, if the sample set $\mathcal{S} \subseteq [N]$ satisfies*

$$|\mathcal{S}| \geq \frac{2}{1 + \sqrt{1 + \frac{2(\sigma_c^{\min})^2}{9(\theta_J + \nu_J)(\sigma_c^{\max})^2}}} N$$

and let $y_{\mathcal{S}}(x) = -\nabla c_{\mathcal{S}}(x)^\dagger \nabla f(x)$, we have

$$\|y_{[N]}(x) - y_{\mathcal{S}}(x)\|_2 \leq \frac{3\sigma_f^{\max}\sigma_c^{\max}}{2(\sigma_c^{\min})^2} \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}}.$$

Proof. By the choice of \mathcal{S} , we have $\sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \leq \frac{\sigma_c^{\min}}{3\sigma_c^{\max}} < \frac{\sigma_c^{\min}}{\sigma_c^{\max}}$ where $\sigma_c^{\min} > 0$, which satisfies requirements of Lemma 3.4. Hence, we have that $\nabla c(x)$ and $\nabla c_{\mathcal{S}}(x)$ are of full column rank and are acute perturbations to each other. By [2, Theorem 5.2], we have the following:

$$(3.7) \quad \|y_{[N]}(x) - y_{\mathcal{S}}(x)\|_2 \leq \underbrace{\frac{\|\nabla c(x)^\dagger\|_2 \|\nabla c(x) - \nabla c_{\mathcal{S}}(x)\|_2}{1 - \|\nabla c(x)^\dagger\|_2 \|\nabla c(x) - \nabla c_{\mathcal{S}}(x)\|_2}}_{(iii)} \|y_{[N]}(x)\|_2.$$

By inequality (3.5) and the choice of \mathcal{S} , we have

$$\|\nabla c_{\mathcal{S}}(x) - \nabla c(x)\|_2 \leq \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \|\nabla c(x)\|_2 \leq \frac{1}{3}\sigma_c^{\min},$$

224 which further gives us

$$225 \quad (3.8) \quad 1 - \|\nabla c(x)^\dagger\|_2 \|\nabla c(x) - \nabla c_{\mathcal{S}}(x)\|_2 \geq 1 - \frac{1}{\sigma_c^{\min}} \frac{\sigma_c^{\min}}{3} = 2/3.$$

226 Hence, we have

$$227 \quad (3.9) \quad \begin{aligned} (iii) &\leq \frac{3}{2} \|\nabla c(x)^\dagger\|_2 \|\nabla c(x) - \nabla c_{\mathcal{S}}(x)\|_2 \\ &\leq \frac{3\sigma_c^{\max}}{2\sigma_c^{\min}} \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}}. \end{aligned}$$

228 Here, the first line uses (3.8) at the denominator, and the last line uses the bound for
229 $\|\nabla c(x)^\dagger\|_2$. Combining with the bound $\|y_{[N]}(x)\|_2 \leq \frac{\sigma_f^{\max}}{\sigma_c^{\min}}$, we have

$$230 \quad \|y_{[N]}(x) - y_{\mathcal{S}}(x)\|_2 \leq \frac{3\sigma_f^{\max}\sigma_c^{\max}}{2(\sigma_c^{\min})^2} \sqrt{\frac{2(\theta_J + \nu_J)N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}}. \quad \square$$

231 The difference in the Hessian conditions is more complicated compared to the
232 gradient condition. Recall the Hessian condition in Definition 3.1 that

$$233 \quad |d^T \nabla_{xx}^2 L(x, y) d| \geq \beta \|d\|_2^2, \quad \forall d \in \text{Null}(\nabla c(x)^T).$$

234 When we consider the empirical system (2.3), not only did the Lagrangian function
235 L change, but also the null space $\text{Null}(\nabla c(x)^T)$ change. We start by giving a general
236 result for two perturbed null spaces by examining the difference between vectors in
237 one null space and their projections onto the other null space.

238 **LEMMA 3.6.** *Under Assumption 3.2 where constants $(\sigma_c^{\min}, \sigma_c^{\max})$ exist. In addi-*
239 *tion, under Assumption 3.3 where constants (θ_J, ν_J) exist. For any $x \in \mathbb{R}^n$ and any*
240 *$\mathcal{S} \subseteq [N]$ such that*

$$241 \quad (3.10) \quad |\mathcal{S}| > \frac{2}{1 + \sqrt{1 + \frac{2(\sigma_c^{\min})^2}{(\theta_J + \nu_J)(\sigma_c^{\max})^2}}} N.$$

242 Then, for any $d_{\mathcal{S}} \in \text{Null}(\nabla c_{\mathcal{S}}(x)^T)$, we have

$$243 \quad \frac{\|\mathcal{R}(\nabla c(x))(d_{\mathcal{S}})\|_2}{\|d_{\mathcal{S}}\|_2} \leq \frac{\sigma_c^{\max}}{\sigma_c^{\min}} \sqrt{\frac{2N(N - |\mathcal{S}|)(\theta_J + \nu_J)}{|\mathcal{S}|^2}} < 1.$$

245 *Proof.* Since $d_{\mathcal{S}} \in \text{Null}(\nabla c_{\mathcal{S}}(x)^T)$ and $\mathcal{N}(\nabla c_{\mathcal{S}}(x))$ is a projection matrix to the
246 null space $\text{Null}(\nabla c_{\mathcal{S}}(x)^T)$, we have $\mathcal{N}(\nabla c_{\mathcal{S}}(x))d_{\mathcal{S}} = d_{\mathcal{S}}$. In addition, by Lemma
247 3.4, (3.10) ensures $\nabla c_{\mathcal{S}}(x)$ and $\nabla c(x)$ are of full column rank. Combining with [2,
248 Theorem 2.4] we have

$$249 \quad (3.11) \quad \begin{aligned} \|\mathcal{R}(\nabla c(x))d_{\mathcal{S}}\|_2 &= \|\mathcal{R}(\nabla c(x))\mathcal{N}(\nabla c_{\mathcal{S}}(x))d\|_2 \\ &\leq \|\nabla c(x)^\dagger\|_2 \|\nabla c(x) - \nabla c_{\mathcal{S}}(x)\|_2 \|d_{\mathcal{S}}\|_2. \end{aligned}$$

250 Then, combined with Lemma 3.3 gives the desired result. Moreover, the choice of $|\mathcal{S}|$
251 (3.10) gives us

$$252 \quad \frac{\sigma_c^{\max}}{\sigma_c^{\min}} \sqrt{\frac{2N(N - |\mathcal{S}|)(\theta_J + \nu_J)}{|\mathcal{S}|^2}} < 1. \quad \square$$

With this result, we can now look into the Hessian condition for empirical constraint Morse problem. In particular, let $d \in \text{Null}(\nabla c_S(x)^T)$, vector $\tilde{d} = \mathcal{N}(\nabla c(x))$ and $r = d - \tilde{d}$, we tend to look at the difference

$$\left| d^T \nabla_{xx}^2 L_S(x, y_S) d - \tilde{d}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) \tilde{d} \right|,$$

and the following lemma gives us a general bound for the above term.

In summary, we have the result for the Morse property of the empirical problem. We define the following three parameters

$$\begin{cases} \eta_1 := \frac{\sigma_c^{\max}}{\sigma_c^{\min}} \sqrt{2(\theta_J + \nu_J)}, \\ \eta_2 := \frac{\sigma_f^{\max} \lambda_c^{\max}}{\sigma_c^{\min}} \sqrt{m \mu_H}, \\ \eta_3 := \eta_2 + 3\eta_1 \lambda_f^{\max} + \frac{9\eta_1 \eta_2}{2\sqrt{\mu_H}}. \end{cases}$$

THEOREM 3.7. *Under Assumption.3.2 and Assumption.3.4 where the constants $(\sigma_c^{\min}, \sigma_c^{\max}, \sigma_f^{\max}, \lambda_c^{\max}, \lambda_f^{\max}, \theta_J, \nu_J, \mu_H)$ exist, and in addition, assuming the problem (2.2) is (α, β) -morse with the dual variable $y_{[N]}$ and y_S are chosen as in (2.6). Then, for any $\mathcal{S} \subseteq [N]$ when satisfies:*

$$(3.12) \quad g_S := \sqrt{\frac{N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}} \leq \min \left\{ \frac{1}{3\eta_1}, \frac{\alpha}{2\sigma_f^{\max}\eta_1}, \frac{\beta}{2\sqrt{(\eta_1\beta + \eta_3)^2 + 3\eta_1\eta_2\beta}} \right\},$$

the problem (2.3) is (α_S, β_S) -morse, where

$$\begin{cases} \alpha_S = \alpha - \sigma_f^{\max} \eta_1 g_S > 0 \text{ and} \\ \beta_S = \beta - (\eta_1\beta + \eta_3) g_S - \frac{3}{2} \eta_1 \eta_2 g_S^2 > 0. \end{cases}$$

Proof. For simplicity of analysis, let $g_S := \sqrt{\frac{N(N - |\mathcal{S}|)}{|\mathcal{S}|^2}}$ when $|\mathcal{S}| \in (0, N]$. By inequality (3.5) and triangle inequality, we have that for any $x \in \mathbb{R}^n$

$$\|\nabla c_S(x)\|_2 \leq \|\nabla c(x)\|_2 + \|\nabla c_S(x) - \nabla c(x)\|_2 \leq \left(1 + \sqrt{2(\theta_J + \nu_J)} g_S\right) \|\nabla c(x)\|_2.$$

Next, we look into the difference between $\nabla_x L_{[N]}(x, y_{[N]})$ and $\nabla_x L_S(x, y_S)$. We have

$$\begin{aligned} & \|\nabla_x L_{[N]}(x, y_{[N]}) - \nabla_x L_S(x, y_S)\|_2 \\ &= \|\nabla c(x) y_{[N]} - \nabla c_S(x) y_S\|_2 \\ (3.13) \quad &= \|\nabla c(x) \nabla c(x)^\dagger \nabla f(x) + \nabla c_S(x) \nabla c_S(x)^\dagger \nabla f(x)\|_2 \\ &\leq \|\nabla c_S(x) + \mathcal{R}(\nabla c(x))\|_2 \|\nabla f(x)\|_2 \\ &\leq \|\nabla c_S(x) + \mathcal{R}(\nabla c(x))\|_2 \sigma_f^{\max}. \end{aligned}$$

Here, the third line uses the definition for $y_{[N]}$ and y_S . The second to last line uses the definition of \mathcal{R} , and the last line uses the bound for $\|\nabla f(x)\|_2$. By choice of \mathcal{S} (3.12), the requirement of Lemma 3.4 is satisfied, and both $\nabla c(x)$ and $\nabla c_S(x)$ are of full column rank. By [2, Theorem 2.4] and previous bounds, we have

$$\begin{aligned} & \|\nabla c_S(x) + \mathcal{R}(\nabla c(x))\|_2 \\ (3.14) \quad &= \|\nabla c_S(x) (I_n - \mathcal{R}(\nabla c(x)))\|_2 \\ &= \|\mathcal{R}(\nabla c_S(x)) \mathcal{N}(\nabla c(x))\|_2 \leq \frac{\sigma_c^{\max}}{\sigma_c^{\min}} \sqrt{2(\theta_J + \nu_J)} g_S = \eta_1 g_S. \end{aligned}$$

Combining this result with the triangle inequality, we have

$$\begin{aligned}\|\nabla_x L_{[N]}(x, y_{[N]})\|_2 &\leq \|\nabla_x L_{\mathcal{S}}(x, y_{\mathcal{S}})\|_2 + \|\nabla_x L_{[N]}(x, y_{[N]}) - \nabla_x L_{\mathcal{S}}(x, y_{\mathcal{S}})\|_2 \\ &\leq \|\nabla_x L_{\mathcal{S}}(x, y_{\mathcal{S}})\|_2 + \sigma_f^{\max} \eta_1 g_{\mathcal{S}}.\end{aligned}$$

Hence for any $x \in \mathbb{R}^n$ satisfies $\|\nabla_x L_{\mathcal{S}}(x, y_{\mathcal{S}})\|_2 \leq \alpha - \sigma_f^{\max} \eta_1 g_{\mathcal{S}} = \alpha_{\mathcal{S}}$, we have $\|\nabla_x L_{[N]}(x, y_{[N]})\|_2 \leq \alpha$. In addition, the choice of \mathcal{S} gives us that $\sigma_f^{\max} \eta_1 g_{\mathcal{S}} \leq \frac{1}{2}\alpha$, we have

$$\alpha_{\mathcal{S}} \geq \frac{1}{2}\alpha > 0.$$

Since the problem (2.2) is (α, β) -morse, by the definition of morse we have

$$|d^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) d| \geq \beta \|d\|_2^2 \text{ for all } d \in \text{Null}(\nabla c(x)^T).$$

Now, for any $d_{\mathcal{S}} \in \text{Null}(\nabla c_{\mathcal{S}}(x)^T)$, we look into the value $|d_{\mathcal{S}}^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y_{\mathcal{S}}) d_{\mathcal{S}}|$. We have

$$\begin{aligned}&|d_{\mathcal{S}}^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y_{\mathcal{S}}) d_{\mathcal{S}}| \\ &= |d_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) d_{\mathcal{S}} + d_{\mathcal{S}}^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y_{\mathcal{S}}) d_{\mathcal{S}} - d_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) d_{\mathcal{S}}| \\ &\geq \underbrace{|d_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) d_{\mathcal{S}}|}_{(v.1)} - \underbrace{|d_{\mathcal{S}}^T \nabla_{xx}^2 L_{\mathcal{S}}(x, y_{\mathcal{S}}) d_{\mathcal{S}} - d_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) d_{\mathcal{S}}|}_{(v.2)}.\end{aligned}$$

Here, we get the third line by adding and subtracting a term and using the triangle inequality.

Let $\tilde{d}_{\mathcal{S}} := \mathcal{N}(\nabla c(x)) d_{\mathcal{S}}$ and $r_{\mathcal{S}} := d_{\mathcal{S}} - \tilde{d}_{\mathcal{S}}$, and substitue $d_{\mathcal{S}} = \tilde{d}_{\mathcal{S}} + r_{\mathcal{S}}$ for term (v.1) we have

$$\begin{aligned}(v.1) &= \left| \tilde{d}_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) \tilde{d}_{\mathcal{S}} + 2 \tilde{d}_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) r_{\mathcal{S}} + r_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) r_{\mathcal{S}} \right| \\ &\geq \left| \tilde{d}_{\mathcal{S}}^T \nabla_{xx}^2 L_{[N]}(x, y_{[N]}) \tilde{d}_{\mathcal{S}} \right| - 2 \|\nabla_{xx}^2 L_{[N]}(x, y_{[N]})\|_2 \|\tilde{d}_{\mathcal{S}}\|_2 \|r_{\mathcal{S}}\|_2 \\ &\quad - \|\nabla_{xx}^2 L_{[N]}(x, y_{[N]})\|_2 \|r_{\mathcal{S}}\|_2^2 \\ &\geq \beta \|\tilde{d}_{\mathcal{S}}\|_2^2 - 3 \|\nabla_{xx}^2 L_{[N]}(x, y_{[N]})\|_2 \|d_{\mathcal{S}}\|_2 \|r_{\mathcal{S}}\|_2.\end{aligned}$$

Here, the first equality and inequality follow by adding, subtracting a term, and using the triangle inequality. The second inequality uses the fact that $\|d_{\mathcal{S}}\|_2 = \|\tilde{d}_{\mathcal{S}}\|_2 + \|r_{\mathcal{S}}\|_2$ which gives $\|\tilde{d}_{\mathcal{S}}\|_2 \leq \|d_{\mathcal{S}}\|_2$, and substitute this result with the last two terms. Further we have

$$\begin{aligned}(v.1) &\geq \beta \|\tilde{d}_{\mathcal{S}}\|_2^2 - 3 \left(\lambda_f^{\max} + \frac{\sqrt{m} \sigma_f^{\max}}{\sigma_c^{\min} \lambda_c^{\max}} g_{\mathcal{S}_k} \right) \|d_{\mathcal{S}}\|_2 \|r_{\mathcal{S}}\|_2 \\ &\geq \beta (1 - \eta_1 g_{\mathcal{S}}) \|d_{\mathcal{S}}\|_2^2 - 3 \left(\lambda_f^{\max} + \frac{\eta_2}{\sqrt{\mu_H}} g_{\mathcal{S}} \right) \eta_1 g_{\mathcal{S}} \|d_{\mathcal{S}}\|_2^2 \\ &= \left(\beta - \left(\eta_1 \beta + 3 \eta_1 \lambda_f^{\max} + 3 \frac{\eta_1 \eta_2}{\sqrt{\mu_H}} g_{\mathcal{S}} \right) g_{\mathcal{S}} \right) \|d_{\mathcal{S}}\|_2^2.\end{aligned}$$

Here, the first line uses Lemma 3.3, the second line uses Lemma 3.6, and the last line rearranges terms.

For the term (v.2), we have

$$\begin{aligned}
(v.2) &= |d_S^T (\nabla_{xx}^2 L_S(x, y_S) - \nabla_{xx}^2 L_{[N]}(x, y_{[N]})) d_S| \\
&= \left| d_S^T \left(\sum_{j=1}^m y_S^j \nabla^2 c_S^j(x) - \sum_{j=1}^m y_{[N]}^j \nabla^2 c^j(x) \right) d_S \right| \\
&\leq \left\| \sum_{j=1}^m (y_S^j \nabla^2 c_S^j(x) - y_{[N]}^j \nabla^2 c^j(x)) \right\|_2 \|d_S\|_2^2,
\end{aligned}
\tag{3.16}$$

where for the term of Hessian, we have

$$\begin{aligned}
&\left\| \sum_{j=1}^m (y_S^j \nabla^2 c_S^j(x) - y_{[N]}^j \nabla^2 c^j(x)) \right\|_2 \\
&= \left\| \sum_{j=1}^m \left(y_S^j (\nabla^2 c_S^j(x) - \nabla^2 c^j(x)) \right) + \sum_{j=1}^m \left((y_S^j - y_{[N]}^j) \nabla^2 c^j(x) \right) \right\|_2 \\
&\leq \sum_{j=1}^m \|y_S^j\|_2 \|\nabla^2 c_S^j(x) - \nabla^2 c^j(x)\|_2 + \sum_{j=1}^m \|y_S^j - y_{[N]}^j\|_2 \|\nabla^2 c^j(x)\|_2 \\
&\leq \sqrt{m\mu_H} \lambda_c^{\max} g_S \|y_S\|_2 + \sqrt{m} \lambda_c^{\max} \|y_S - y_{[N]}\|_2 \\
&\leq \sqrt{m\mu_H} \lambda_c^{\max} g_S \|y_{[N]}\|_2 + (1 + \sqrt{\mu_H} g_S) \sqrt{m} \lambda_c^{\max} \|y_S - y_{[N]}\|_2 \\
&\leq \sqrt{m\mu_H} \frac{\lambda_c^{\max} \sigma_f^{\max}}{\sigma_c^{\min}} g_S + (1 + \sqrt{\mu_H} g_S) \sqrt{m} \frac{3\sigma_f^{\max} \lambda_c^{\max} \sigma_c^{\max}}{2(\sigma_c^{\min})^2} \sqrt{2(\theta_J + \nu_J)} g_S \\
&= \eta_2 g_S + \frac{3\eta_1 \eta_2}{2\sqrt{\mu_H}} g_S + \frac{3\eta_1 \eta_2}{2} g_S^2.
\end{aligned}$$

Here, the second is to add, subtract, and rearrange terms. The third line uses triangle inequality and submultiplicity. The fourth line uses similar arguments as in (v.1). The fifth line uses the fact that $\|y_S\|_2 \leq \|y_{[N]}\|_2 + \|y_S - y_{[N]}\|_2$ and rearranges terms. The last two lines use Lemma 3.5 since $g_S \leq \frac{1}{3\eta_1}$, and the definition of (η_1, η_2) .

Combining the above results for (v.1, v.2), we have

$$\begin{aligned}
&|d_S^T \nabla_{xx}^2 L_S(x, y_S) d_S| \\
&\geq \left(\beta - \left(\eta_1 \beta + \eta_2 + 3\eta_1 \lambda_f^{\max} + \frac{9\eta_1 \eta_2}{2\sqrt{\mu_H}} \right) g_S - \frac{3}{2} \eta_1 \eta_2 g_S^2 \right) \|d_S\|_2^2 \\
&= \left(\beta - (\eta_1 \beta + \eta_3) g_S - \frac{3}{2} \eta_1 \eta_2 g_S^2 \right) \|d_S\|_2^2.
\end{aligned}$$

By the requirement of $|\mathcal{S}|$ that, we have

$$\left((\eta_1 \beta + \eta_3) g_S + \frac{3}{2} \eta_1 \eta_2 g_S^2 \right) \leq \frac{1}{2} \beta,$$

where the nonnegative solution for g_S is

$$0 \leq g_S \leq \frac{-(\eta_1 \beta + \eta_3) + \sqrt{(\eta_1 \beta + \eta_3)^2 + 3\eta_1 \eta_2 \beta}}{3\eta_1 \eta_2},$$

where the right-hand side can be bounded below by

$$\begin{aligned}
& \frac{-(\eta_1\beta + \eta_3) + \sqrt{(\eta_1\beta + \eta_3)^2 + 3\eta_1\eta_2\beta}}{3\eta_1\eta_2} \\
&= \frac{3\eta_1\eta_2\beta}{9\eta_1\eta_2 \left((\eta_1\beta + \eta_3) + \sqrt{(\eta_1\beta + \eta_3)^2 + 3\eta_1\eta_2\beta} \right)} \\
&\geq \frac{\beta}{2\sqrt{(\eta_1\beta + \eta_3)^2 + 3\eta_1\eta_2\beta}}.
\end{aligned}$$

Here, the second line multiplies a $\left((\eta_1\beta + \eta_3) + \sqrt{(\eta_1\beta + \eta_3)^2 + 3\eta_1\eta_2\beta} \right)$ at both the numerator and denominator. Hence the last requirement for \mathcal{S} ensures that. \square

THEOREM 3.8. *Under Assumption 3.2 and Assumption 3.4 where the constants $(\sigma_c^{\min}, \sigma_c^{\max}, \sigma_f^{\max}, \lambda_c^{\max}, \lambda_f^{\max}, \theta_J, \nu_J, \mu_H)$ exist, and assume problem (2.2) is (α, β) -morse. Define tolerances*

$$\epsilon_k := \eta_1 \sigma_f^{\max} \sqrt{\frac{N(N - |\mathcal{S}_k|)}{|\mathcal{S}_k|^2}} \text{ and } \varepsilon_k := \eta_1 \lambda_f^{\max} \sqrt{\frac{N(N - |\mathcal{S}_k|)}{|\mathcal{S}_k|^2}}, \forall k \in [K].$$

Let (2.2) proceed with Algorithm 2.1. Then, for all sample sets $\mathcal{S}_k \subseteq [N]$ satisfies

$$(3.17) \quad \sqrt{\frac{N(N - |\mathcal{S}_k|)}{|\mathcal{S}_k|^2}} \leq \min \left\{ \frac{1}{3\eta_1}, \frac{\alpha}{4\eta_1 \sigma_f^{\max}}, \frac{\beta}{4\sqrt{(\eta_1\beta + \eta_3)^2 + \frac{3}{2}\eta_1\eta_2\beta}} \right\},$$

if $x_{\mathcal{S}_k} \in \mathbb{R}^n$ is a $(\epsilon_k, \varepsilon_k)$ stationary solution, the $x_{\mathcal{S}_k}$ must satisfy the following for problem (2.3) with $\mathcal{S} = \mathcal{S}_{k+1}$

$$\begin{aligned}
& \|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, y_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}))\|_2 \leq \alpha_{\mathcal{S}_{k+1}}, \text{ and} \\
& d^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, y_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})) d \geq \beta_{\mathcal{S}_{k+1}} \|d\|_2^2, \forall d \in \text{Null}(\nabla c_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})^T).
\end{aligned}$$

Proof. Let the dual variables $y_{[N]}$ and $y_{\mathcal{S}_k}$ be defined as in (2.6). In addition, define $z_{\mathcal{S}_{k+1}} = -\nabla c_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})^\dagger \nabla f(x_{\mathcal{S}_k})$. Similar to (3.13) we have

$$\begin{aligned}
& \|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) - \nabla_x L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k})\|_2 \\
& \leq \|\mathcal{R}(\nabla c_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})) - \mathcal{R}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k}))\|_2 \|\nabla f(x_{\mathcal{S}_k})\|_2 \\
& \leq (\|\mathcal{R}(\nabla c_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})) - \mathcal{R}(\nabla c(x_{\mathcal{S}_k}))\|_2 \\
& \quad + \|\mathcal{R}(\nabla c(x_{\mathcal{S}_k})) - \mathcal{R}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k}))\|_2) \|\nabla f(x_{\mathcal{S}_k})\|_2.
\end{aligned}$$

Here, the last inequality uses the triangle inequality. In (3.14) we already have

$$\|\mathcal{R}(\nabla c_{\mathcal{S}}(x)) - \mathcal{R}(\nabla c(x))\|_2 \leq \eta_1 g_{\mathcal{S}}.$$

Note the right-hand side depends on $g_{\mathcal{S}}$, which by definition decreases when $|\mathcal{S}|$ increases. Hence we have

$$\|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) - \nabla_x L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k})\|_2 \leq 2\eta_1 \sigma_f^{\max} g_{\mathcal{S}},$$

which further gives us

$$\|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}})\|_2 \leq \|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) - \nabla_x L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k})\|_2$$

$$\begin{aligned}
& + \|\nabla_x L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k})\|_2 \\
& \leq 3\eta_1 \sigma_f^{\max} g_{\mathcal{S}_k} \leq \frac{3}{4}\alpha.
\end{aligned}$$

Here, the first inequality uses the triangle inequality. The second inequality combines with the fact that $x_{\mathcal{S}_k}$ is a $(\epsilon_k, \varepsilon_k)$ stationary point, and the last inequality comes from the first requirement for $|\mathcal{S}|$ that $\eta_1 \sigma_f^{\max} g_{\mathcal{S}_k} \leq \frac{1}{4}\alpha$.

Moreover, the same requirement for $|\mathcal{S}|$ gives us

$$\alpha_{\mathcal{S}_k} = \alpha - \eta_1 \sigma_f^{\max} g_{\mathcal{S}_k} \geq \frac{3}{4}\alpha,$$

and combining with the fact that $\alpha_{\mathcal{S}}$ decreases when $|\mathcal{S}|$ increases, we have

$$(3.18) \quad \|\nabla_x L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}})\|_2 \leq \alpha_{\mathcal{S}_k} \leq \alpha_{\mathcal{S}_{k+1}}.$$

Now, we turn to the condition for hessian. Since the subproblem for \mathcal{S}_{k+1} is $(\alpha_{\mathcal{S}_{k+1}}, \beta_{\mathcal{S}_{k+1}})$ -morse and with (3.18), we have

$$\left| d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} \right| \geq \beta_{\mathcal{S}_{k+1}} \|d_{\mathcal{S}_{k+1}}\|_2^2, \quad \forall d_{\mathcal{S}_{k+1}} \in \text{Null}(\nabla c_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k})^T).$$

Similar to the analysis for Theorem (3.7), define $\bar{d}_{\mathcal{S}_{k+1}} := \mathcal{N}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k})) d_{\mathcal{S}_{k+1}}$, by triangle inequality we have

$$\begin{aligned}
& d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} \\
& \geq \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k}) \bar{d}_{\mathcal{S}_{k+1}} \\
& \quad - \underbrace{\left| d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k}) \bar{d}_{\mathcal{S}_{k+1}} \right|}_{(vi)} \\
& \geq -\varepsilon_k \|d_{\mathcal{S}_{k+1}}\|_2^2 - (vi).
\end{aligned}$$

Here, the last line uses the termination condition (2.5b) and the fact that $\|\bar{d}_{\mathcal{S}_{k+1}}\|_2^2 \leq \|d_{\mathcal{S}_{k+1}}\|_2^2$.

To give a bound for (vi), we add and subtract four terms. Define the variable $z_{[N]} := -\nabla c(x_{\mathcal{S}_k})^\dagger \nabla f(x_{\mathcal{S}_k})$, following the triangle inequality, we have

$$\begin{aligned}
(vi) & = \left| d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} - d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) d_{\mathcal{S}_{k+1}} \right. \\
& \quad + d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) d_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) \bar{d}_{\mathcal{S}_{k+1}} \\
& \quad \left. + \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) \bar{d}_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k}) \bar{d}_{\mathcal{S}_{k+1}} \right| \\
& \leq \underbrace{\left| d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} - d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) d_{\mathcal{S}_{k+1}} \right|}_{(vi.1)} \\
& \quad + \underbrace{\left| d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) d_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) \bar{d}_{\mathcal{S}_{k+1}} \right|}_{(vi.2)} \\
& \quad + \underbrace{\left| \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]}) \bar{d}_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_k}(x_{\mathcal{S}_k}, y_{\mathcal{S}_k}) \bar{d}_{\mathcal{S}_{k+1}} \right|}_{(vi.3)}.
\end{aligned}$$

Thanks to the previous result in Theorem 3.7 on the term (v.2), we have

$$(vi.1) \leq \left(\eta_2 g_{\mathcal{S}_{k+1}} + \frac{3\eta_1\eta_2}{2\sqrt{\mu_H}} g_{\mathcal{S}_{k+1}} + \frac{3\eta_1\eta_2}{2} g_{\mathcal{S}_{k+1}}^2 \right) \|d_{\mathcal{S}_{k+1}}\|_2^2, \text{ and}$$

$$(vi.3) \leq \left(\eta_2 g_{\mathcal{S}_k} + \frac{3\eta_1\eta_2}{2\sqrt{\mu_H}} g_{\mathcal{S}_k} + \frac{3\eta_1\eta_2}{2} g_{\mathcal{S}_k}^2 \right) \|\bar{d}_{\mathcal{S}_{k+1}}\|_2^2.$$

For (vi.2), by Lemma 3.3 we have $\|\nabla_{xx}^2 L_{[N]}(x_{\mathcal{S}_k}, z_{[N]})\|_2 \leq \lambda_f^{\max} + \frac{\eta_2}{\sqrt{\mu_H}}$, and that

$$\begin{aligned} \|d_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}\|_2 &= \|d_{\mathcal{S}_{k+1}} - \mathcal{N}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k}))d_{\mathcal{S}_{k+1}}\|_2 \\ &= \|\mathcal{R}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k}))d_{\mathcal{S}_{k+1}}\|_2 \\ &\leq \|\mathcal{R}(\nabla c(x_{\mathcal{S}_k}))d_{\mathcal{S}_{k+1}}\|_2 + \|(\mathcal{R}(\nabla c(x_{\mathcal{S}_k})) - \mathcal{R}(\nabla c_{\mathcal{S}_k}(x_{\mathcal{S}_k})))d_{\mathcal{S}_{k+1}}\|_2 \\ &\leq 2\eta_1 g_{\mathcal{S}_k} \|d_{\mathcal{S}_{k+1}}\|_2. \end{aligned}$$

Here, the first line uses the definition of $\bar{d}_{\mathcal{S}_{k+1}}$. The second line uses the definition of \mathcal{R} . The third line uses the triangle inequality. The last line uses Lemma 3.6, and inequality (3.14). In addition, the second line also gives us $\|d_{\mathcal{S}_{k+1}} - \bar{d}_{\mathcal{S}_{k+1}}\|_2 \leq \|d_{\mathcal{S}_{k+1}}\|_2$.

Combining all the above results, noticing that $g_{\mathcal{S}}$ is nondecreasing with respect to $|\mathcal{S}|$, which gives $g_{\mathcal{S}_{k+1}} \leq g_{\mathcal{S}_k}$. And remember that $\|\bar{d}_{\mathcal{S}_{k+1}}\|_2 \leq \|d_{\mathcal{S}_{k+1}}\|_2$ and $\varepsilon_k = \eta_1 \lambda_f^{\max} g_{\mathcal{S}_k}$, we have

$$\begin{aligned} &d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} \\ &\geq -\eta_1 \lambda_f^{\max} g_{\mathcal{S}_k} \|d_{\mathcal{S}_{k+1}}\|_2^2 - 2 \left(\eta_2 + \frac{3\eta_1\eta_2}{2\sqrt{\mu_H}} + \frac{3\eta_1\eta_2}{2} g_{\mathcal{S}_k} \right) g_{\mathcal{S}_k} \|d_{\mathcal{S}_{k+1}}\|_2^2 \\ &\quad - \left(\lambda_f^{\max} + \frac{\eta_2}{\sqrt{\mu_H}} \right) 2\eta_1 g_{\mathcal{S}_k} \|d_{\mathcal{S}_{k+1}}\|_2^2 \\ &= - \left(\left(3\eta_1 \lambda_f^{\max} + 2\eta_2 + \frac{5\eta_1\eta_2}{\sqrt{\mu_H}} \right) g_{\mathcal{S}_k} + 3\eta_1\eta_2 g_{\mathcal{S}_k}^2 \right) \|d_{\mathcal{S}_{k+1}}\|_2^2 \\ &\geq - (2\eta_3 g_{\mathcal{S}_k} + 3\eta_1\eta_2 g_{\mathcal{S}_k}^2) \|d_{\mathcal{S}_{k+1}}\|_2^2. \end{aligned}$$

Similar to the analysis for Theorem 3.7. Recall $\beta_{\mathcal{S}} = \beta - ((\eta_1\beta + \eta_3)g_{\mathcal{S}_k} + \frac{3}{2}\eta_1\eta_2 g_{\mathcal{S}_k}^2)$ and $\beta_{\mathcal{S}_{k+1}} \geq \beta_{\mathcal{S}_k}$. To ensure $\beta_{\mathcal{S}_{k+1}} \geq \frac{3}{4}\beta$, we need $\frac{1}{4}\beta \geq (\eta_1\beta + \eta_3)g_{\mathcal{S}_k} + \frac{3}{2}\eta_1\eta_2 g_{\mathcal{S}_k}^2$, whose nonnegative solution is

$$g_{\mathcal{S}_k} \leq \frac{-(\eta_1\beta + \eta_3) + \sqrt{(\eta_1\beta + \eta_3)^2 + \frac{3}{2}\eta_1\eta_2\beta}}{3\eta_1\eta_2},$$

and it is ensured by

$$g_{\mathcal{S}_k} \leq \frac{\beta}{4\sqrt{(\eta_1\beta + \eta_3)^2 + \frac{3}{2}\eta_1\eta_2\beta}}.$$

Moreover, we have

$$-(2\eta_3 g_{\mathcal{S}_k} + 3\eta_1\eta_2 g_{\mathcal{S}_k}^2) \geq -2 \left((\eta_3 + \eta_1\beta) g_{\mathcal{S}_k} - \frac{3}{2}\eta_1\eta_2 g_{\mathcal{S}_k}^2 \right) \geq -\frac{1}{2}\beta > -\beta_{\mathcal{S}_{k+1}}.$$

The above analysis gives us that $d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} > -\beta_{\mathcal{S}_{k+1}} \|d_{\mathcal{S}_{k+1}}\|_2^2$. However, since subproblem (2.3) for $\mathcal{S} = \mathcal{S}_{k+1}$ is $(\alpha_{k+1}, \beta_{k+1})$ -morse, which says that

409 $|d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}}| \geq \beta_{\mathcal{S}_{k+1}} \|d_{\mathcal{S}_{k+1}}\|_2^2$. Combining these two, we must
 410 have

$$411 \quad d_{\mathcal{S}_{k+1}}^T \nabla_{xx}^2 L_{\mathcal{S}_{k+1}}(x_{\mathcal{S}_k}, z_{\mathcal{S}_{k+1}}) d_{\mathcal{S}_{k+1}} \geq \beta_{\mathcal{S}_{k+1}} \|d_{\mathcal{S}_{k+1}}\|_2^2,$$

412 which completes proof. \square

413

414 **ASSUMPTION 3.4.** *We make the following assumptions for each element of the*
 415 *expected constraint function, c^i . There exists a $(r, \tau) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that for all*
 416 *$(x, i) \in \mathbb{B}^n(r) \times \{1, \dots, m\}$,*

417 *(1). the gradient of $c^i(x)$ is τ^2 -sub-Gaussian. Namely, for any $a \in \mathbb{R}^n$,*

$$418 \quad \mathbb{E}_\xi [\exp(a^T (\nabla c^i(x; \xi) - \mathbb{E}_\xi [\nabla c^i(x; \xi)]))] \leq \exp\left(\frac{\tau^2 \|a\|_2^2}{2}\right).$$

419 *z_{ξ_j} finite sample distribution*

420 *(2). the Hessian of $c^i(x)$, evaluated on a unit vector, is τ^2 -sub-exponential. Namely,*
 421 *for any $a \in \mathbb{B}^n(1)$, let $z_{a,x,\xi} := a^T \nabla^2 c^i(x; \xi) a$, then*

$$422 \quad \mathbb{E}_\xi \left[\exp\left(\frac{1}{\tau^2} |z_{a,x,\xi} - \mathbb{E}[z_{a,x,\xi}]| \right) \right] \leq 2.$$

423 *(3). within $\mathbb{B}^n(r)$, the Hessian of c^i is L -Lipschitz continuous, and the gradient*
 424 *of c^i is λ_c^{\max} -Lipschitz continuous. Moreover, there exists a constant $h > 0$*
 425 *such that*

$$426 \quad L \leq \tau^3 n^h, \text{ and } \lambda_c^{\max} \leq \tau^2 n^h.$$

427 **THEOREM 3.9.** *Under Assumption.3.4 and let $(n, r, \tau, h) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$*
 428 *be defined in the same way. There exists a universal constant C_0 and for any $\delta \in [0, 1]$*
 429 *let $C := C_0 \max\{h, \log \frac{r\tau}{\delta}, 1\}$. Then, for any sample size $p \geq Cn \log n$, the following*
 430 *holds with probability at least $(1 - \delta)$:*

$$431 \quad (3.20) \quad \sup_{\forall x \in \mathbb{B}^n(r)} \|\nabla c(x) - \nabla c_p(x)\|_2 \leq g(p) := \tau \sqrt{\frac{Cn \log p}{p}} \text{ and}$$

$$\sup_{i \in \{1, \dots, m\}} \left\{ \sup_{\forall x \in \mathbb{B}^n(r)} \|\nabla^2 c_p^i(x) - \nabla^2 c^i(x)\|_2 \right\} \leq G(p) := \tau^2 \sqrt{\frac{Cn \log p}{p}}.$$

432 **LEMMA 3.10.** *Under Assumption.3.4 and let $(n, r, \tau, h) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$*
 433 *be defined in the same way. Let (C, p) be defined in the same way as Theorem.3.9,*
 434 *then the following holds with probability at least $(1 - \delta)$:*

$$(3.21) \quad \sup_{\forall x \in \mathbb{B}^n(r)} \left\| \frac{1}{p} \sum_{i \in \mathcal{S}_p} \nabla c(x, \xi_i) - \frac{1}{2p} \sum_{i \in \mathcal{S}_{2p}} \nabla c(x, \xi_i) \right\|_2 \leq \tau \sqrt{\frac{Cn \log p}{p}} \text{ and}$$

$$\sup_{i \in \{1, \dots, m\}} \left\{ \sup_{\forall x \in \mathbb{B}^n(r)} \left\| \frac{1}{p} \sum_{i \in \mathcal{S}_p} \nabla^2 c^i(x, \xi_i) - \frac{1}{2p} \sum_{i \in \mathcal{S}_{2p}} \nabla^2 c^i(x, \xi_i) \right\|_{op} \right\} \leq \tau^2 \sqrt{\frac{Cn \log p}{p}}.$$

436 *Proof.* \square

437 **4. Numerical Results.**

438 **5. Conclusion.**

439 **6. Acknowledgments.** This work was supported by Office of Naval Research
440 award N00014-24-1-2703.

441 REFERENCES

- 442 [1] A. MOKHTARI, A. OZDAGLAR, AND A. JADBABAIE, *Efficient nonconvex empirical risk minimiza-*
443 *tion via adaptive sample size methods*, in The 22nd International Conference on Artificial
444 Intelligence and Statistics, PMLR, 2019, pp. 2485–2494.
445 [2] G. W. STEWART, *On the perturbation of pseudo-inverses, projections and linear least squares*
446 *problems*, SIAM Review, 19 (1977), pp. 634–662, <http://www.jstor.org/stable/2030248> (ac-
447 cessed 2024-10-21).
448 [3] G. W. STEWART, *Perturbation theory for the singular value decomposition*, Citeseer, 1998.