Appendix of
“Derivative-Free Optimization via Classification”

Yang Yu and Hong Qian and Yi-Qi Hu
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{yuy,qianh,huyq}@lamda.nju.edu.cn

In this appendix, we first introduce some definitions and notations in Section 1. Then, we prove Lemma 1 in Section 2 and Lemma 2 in Section 3. Theorem 1 is proved in Section 4, and the proofs of Corollary 1, 2 and 3 are presented in Section 5, 6 and 7, respectively.

Definitions and Notations

Let $X$ denote a solution space that is a compact subset of $\mathbb{R}^n$, and $f: X \rightarrow \mathbb{R}$ denote a minimization problem. Assume that there exist $x^*, x' \in X$ such that $f(x^*) = \min_{x \in X} f(x)$ and $f(x') = \max_{x \in X} f(x)$. Let $\mathcal{F}$ denote a collection of functions that satisfy this assumption. Given $f \in \mathcal{F}$, the minimization problem is to find a solution $x^* \in X$ s.t. $f(x^*) \leq f(x)$ for all $x \in X$.

For a subset $D \subseteq X$, let $\#D = \int_{x \in X} \mathbb{1}[x \in D] \, dx$ (or $\#D = \sum_{x \in X} \mathbb{1}[x \in D]$ for finite discrete domains), where $\mathbb{1}[]$ is the indicator function. Define $|D| = \#D/\#X$ and thus $|D| \in [0,1]$. Let $D_\alpha = \{x \in X \mid f(x) \leq \alpha\}$, and $D_\epsilon = \{x \in X \mid f(x) - f(x^*) \leq \epsilon\}$ for $\epsilon > 0$. Let $\Delta$ denote the symmetric difference of two sets defined as $A_1 \Delta A_2 = (A_1 \cup A_2) - (A_1 \cap A_2)$. A hypothesis is a mapping $h: X \rightarrow \{-1, +1\}$. Let $\mathcal{H} \subseteq \{h: X \rightarrow \{-1, +1\}\}$ be a hypothesis space. Let $D_h = \{x \in X \mid h(x) = +1\}$ for hypothesis $h \in \mathcal{H}$, i.e., the positive class region represented by $h$. Denote $\mathcal{U}_X$ and $\mathcal{U}_{D_h}$ the uniform distribution over $X$ and $D_h$, respectively, and denote $\mathcal{T}_{D_h}$ the distribution defined on $D_h$ induced by $h$. Let $D_{KL}$ denote the Kullback-Leibler (KL) divergence between two probability distributions. Let $\log(\cdot)$ and $\ln(\cdot)$ be the base two logarithm and natural logarithm, respectively. Let $\text{poly}(\cdot)$ be the set of all polynomials w.r.t. the related variables and $\text{superpoly}(\cdot)$ be the set of all functions that grow faster than any function in $\text{poly}(\cdot)$ w.r.t. the related variables.

Proof of Lemma 1

**Lemma 1.** Given $f \in \mathcal{F}$, $0 < \delta < 1$ and $\epsilon > 0$, the $(\epsilon,\delta)$-query complexity of a classification-based optimization algorithm is upper bounded by

$$O \left( \max \left\{ \frac{1}{(1 - \lambda)|D_x| + \lambda \mathbf{Pr}_h}, \ln \frac{1}{\delta \cdot \sum_{t=1}^{T} m_{\mathbf{Pr}_{h_t}}} \right\} \right),$$

where $\mathbf{Pr}_h = \frac{1}{T} \sum_{t=1}^{T} \mathbf{Pr}_{h_t} = \frac{1}{T} \sum_{t=1}^{T} \int_{x \in D_h} \mathcal{U}_{D_h}(x) \, dx$ (or $\mathbf{Pr}_h = \frac{1}{T} \sum_{t=1}^{T} \sum_{x \in D_h} \mathcal{U}_{D_h}(x)$ for finite discrete domains) is the average success probability of sampling from the learned positive area of $h_t$.

**Proof.** In each iteration, $m_{\mathbf{Pr}_{h_t}}$ samples are needed to realize the probability $\mathbf{Pr}_{h_t}$. Generally speaking, the higher the probability the larger the sample size, but it depends on the concrete implementation of the algorithm. Thus, $\sum_{t=1}^{T} m_{\mathbf{Pr}_{h_t}}$ number of samples is naturally required. We next prove the rest of the bound.
The total number of calls to \( \mathcal{O} \) by a classifier-based optimization algorithm is \((m+1)T\). We consider the probability that, after \( T \) iterations, a classifier-based optimization algorithm outputs a bad solution \( \hat{x} \) s.t. \( f(\hat{x}) - f(x^*) > \epsilon \), and we denote it as \( \Pr(f(\hat{x}) - f(x^*) > \epsilon) \). Since \( \hat{x} \) is the best solution among all sampled solutions, \( \Pr(f(\hat{x}) - f(x^*) > \epsilon) \) is the probability of intersection of events that sampling in each step does not generate a good solution \( x \) s.t. \( f(x) - f(x^*) \leq \epsilon \). For sampling from \( \mathcal{U}_X \), the probability of failure is \( 1 - \Pr_{u} \), where \( \Pr_{u} = \#D_{\epsilon} / \#X \) is the success probability of uniform sampling in \( X \). For sampling from the distribution \( \mathcal{T}_{h_t} \) defined on \( D_{h_t} \) induced by the learned hypothesis \( h_t \), the probability of failure is \( 1 - \Pr_{h_t} \), where \( \Pr_{h_t} = \int_{D_{h_t}} \mathcal{T}_{h_t}(x) \, dx \) (or \( \Pr_{h_t} = \sum_{x \in D_{h_t}} \mathcal{T}_{h_t}(x) \) for finite discrete domains) is the success probability of sampling from \( \mathcal{T}_{h_t} \). Let \( \exp(x) \) denote \( e^x \). Since that every sampling is independent, we have

\[
\Pr(f(\hat{x}) - f(x^*) > \epsilon) = (1 - \Pr_{u})^m \prod_{t=1}^{T} \left( \frac{1 - \Pr_{u}}{1 - \Pr_{h_t}} \right)^m \leq \exp \left( - (1 - x) \Pr_{u} \cdot m + \lambda \Pr_{h_t} \cdot m \right)
\]

In order to let \( \Pr(f(\hat{x}) - f(x^*) > \epsilon) < \delta \), it suffices that

\[
\exp \left( - (1 - \lambda) \Pr_{u} + \lambda \Pr_{h_t} \right) \cdot mT < \delta.
\]

Therefore, we derive that \( mT \in O \left( \frac{1}{(1 - \lambda) \Pr_{u} + \lambda \Pr_{h_t}} \ln \frac{1}{\delta} \right) \). At last, by \( m + 1 \) iterations, we prove the lemma.

**Proof of Lemma 2**

Let \( R_D \) denote the generalization error of \( h \in \mathcal{H} \) with respect to the target function under distribution \( D \), and \( D_{KL} \) denote the Kullback-Leibler (KL) divergence between two probability distributions.

**Lemma 2**

Given \( f \in \mathcal{F} \), \( \epsilon > 0 \), the average success probability of sampling from the distributions induced by the learned hypotheses of any classifier-based optimization algorithm \( \Pr_{\hat{h}_t} \) is lower bounded by

\[
\Pr_{\hat{h}_t} \geq \frac{1}{T} \sum_{t=1}^{T} \left( |D_{\alpha_t}| - 2 \Psi_{D_{KL}(D_t || \mathcal{U}_X)}^{R_{D_t}} \right) / \left( |D_{\alpha_t}| + \Psi_{D_{KL}(D_t || \mathcal{U}_X)}^{R_{D_t}} \right),
\]

where \( D_{\epsilon} = \lambda \mathcal{U}_{D_{h_t}} + (1 - \lambda) \mathcal{U}_X \) is the sampling distribution at iteration \( t \), and \( \Psi_{D_{KL}(D_t || \mathcal{U}_X)}^{R_{D_t}} = R_{D_t} + \#X \sqrt{\frac{1}{2} D_{KL}(D_t || \mathcal{U}_X)} \).
To prove this lemma, our strategy is to first bound $\Pr_{h_t}$, which is the success probability of sampling from the distributions induced by the learned hypothesis at iteration $t$, and then bound $\Pr_h$ by definition.

Bounding $\Pr_{h_t}$

In this section, we will bound $\Pr_{h_t}$ by two steps. A primary lower bound of $\Pr_{h_t}$ is shown in Lemma 3 below, and an explicit lower bound will be presented later.

**Lemma 3**

Given $f \in \mathcal{F}$, $\epsilon > 0$ and any hypothesis $h_t \in \mathcal{H}$, $\Pr_{h_t}$ is lower bounded by

$$\Pr_{h_t} \geq \frac{|D_e \cap D_{h_t}|}{|D_{h_t}|} - \frac{1}{2} D_{KL}(\mathcal{T}_{h_t} \| |U_{D_{h_t}}|),$$

where $D_{KL}$ denotes the Kullback-Leibler (KL) divergence between two probability distributions.

**Proof.** We only consider continuous domains situation and omit finite discrete domains situation since the proof procedure is quite similar. Let $\mathbb{I}[:]$ denote the indicator function, the proof starts from the definition of $\Pr_{h_t}$.

$$\Pr_{h_t} = \int_{D_{h_t}} \mathcal{T}_{h_t}(x) \cdot \mathbb{I}[x \in D_e] \, dx = \int_{D_{h_t}} (\mathcal{T}_{h_t}(x) - U_{D_{h_t}}(x) + U_{D_{h_t}}(x)) \cdot \mathbb{I}[x \in D_e] \, dx$$

$$= \frac{|D_e \cap D_{h_t}|}{|D_{h_t}|} + \int_{D_{h_t}} (\mathcal{T}_{h_t}(x) - U_{D_{h_t}}(x)) \cdot \mathbb{I}[x \in D_e] \, dx$$

$$\geq \frac{|D_e \cap D_{h_t}|}{|D_{h_t}|} - \frac{1}{2} D_{KL}(\mathcal{T}_{h_t} \| |U_{D_{h_t}}|) \int_{D_{h_t}} \mathbb{I}[x \in D_e] \, dx$$

where $U_{D_{h_t}}$ is the uniform distribution over $D_{h_t}$, and the last inequality is by the Pinsker’s inequality. ■

In order to derive an more explicit lower bound of $\Pr_{h_t}$, we need to investigate $|D_{h_t}|$ and $|D_e \cap D_{h_t}|$, and we will bound them respectively.

Bounding $|D_{h_t}|$

**Lemma 4**

Given $f \in \mathcal{F}$ and any hypothesis $h_t \in \mathcal{H}$, $|D_{h_t}|$ is bounded by

$$|D_{\alpha_t}| - R_{t\mathcal{X},t} \leq |D_{h_t}| \leq |D_{\alpha_t}| + R_{t\mathcal{X},t},$$

where $R_{t\mathcal{X},t}$ is the generalization error of $h_t$ with respect to $D_{\alpha_t}$ under distribution $\mathcal{U}_X$.

**Proof.** Let $\Delta$ denote the symmetric difference operator of two sets. We can verify directly that $|\mathcal{D}_{h_t} \setminus |D_{\alpha}| \leq |D_{\alpha_t} \Delta D_{\alpha_t}| = R_{t\mathcal{X},t}$, where $R_{t\mathcal{X},t}$ is the generalization error of $h_t$ with respect to $D_{\alpha_t}$ under distribution $\mathcal{U}_X$. Thus, $|D_{\alpha_t}| - R_{t\mathcal{X},t} \leq |D_{h_t}| \leq |D_{\alpha_t}| + R_{t\mathcal{X},t}$. ■

Bounding $|D_e \cap D_{h_t}|$

**Lemma 5**

Given $f \in \mathcal{F}$, $\epsilon > 0$ and any hypothesis $h_t \in \mathcal{H}$, $|D_e \cap D_{h_t}|$ is lower bounded by

$$|D_e \cap D_{h_t}| \geq |D_e| - 2R_{t\mathcal{X},t},$$

where $R_{t\mathcal{X},t}$ is the generalization error of $h_t$ with respect to $D_{\alpha_t}$ under distribution $\mathcal{U}_X$.  

3
**Proof.** We w.l.o.g. assume that \( \epsilon \leq \alpha_\epsilon \) for all \( t \). Let \( \Delta \) denote the symmetric difference operator of two sets, by set operators, we have

\[
|D_\epsilon \cap D_{h_t}| = |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{h_t}|
\]

\[
\geq |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{\alpha_\epsilon}| - |D_{\alpha_\epsilon} \Delta D_{h_t}|
\]

\[
= |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{\alpha_\epsilon}| - R_{\Delta,t}
\]

\[
= |D_\epsilon \cup D_{h_t}| + |D_\epsilon| - |D_{\alpha_\epsilon}| - R_{\Delta,t}
\]

\[
\geq |D_{h_t}| + |D_\epsilon| - |D_{\alpha_\epsilon}| - R_{\Delta,t},
\]

where the first inequality is by the triangle inequality, and the last equality is by \( D_\epsilon \subseteq D_{\alpha_\epsilon} \). Combining it with the conclusion of Lemma 4 results in that

\[
|D_\epsilon \cap D_{h_t}| \geq (|D_{h_t}| - |D_{\alpha_\epsilon}|) + |D_\epsilon| - R_{\Delta,t} \geq |D_\epsilon| - 2R_{\Delta,t}.
\]

**Bounding \( |D_{h_t}| \) and \( |D_\epsilon \cap D_{h_t}| \) More Explicitly**

Lemma 4 and 5 show that \( |D_{h_t}| \) and \( |D_\epsilon \cap D_{h_t}| \) are bounded by the generalization error \( R_{\Delta,t} \) of \( h_t \) under \( \mathcal{U}_X \). Since the true sampling distribution in the classifier-based optimization framework at each iteration is \( \mathcal{D}_t = \lambda \mathcal{T}_{h_t} + (1 - \lambda) \mathcal{U}_X \) instead of \( \mathcal{U}_X \), it is necessary to investigate the relationship between \( R_{\Delta,t} \) and \( R_{\Delta} \) in order to bound \( |D_{h_t}| \) and \( |D_\epsilon \cap D_{h_t}| \) more explicitly via \( R_{\Delta} \).

**Lemma 6**

The generalization error \( R_{\Delta,t} \) of \( h \) under \( \mathcal{U}_X \) and the generalization error \( R_{\Delta} \) of \( h \) under any distribution \( \mathcal{D} \) have the following relationship:

\[
R_{\Delta,t} \leq R_{\Delta} + \#X \sqrt{\frac{1}{2} D_{KL}(\mathcal{D}||\mathcal{U}_X)}.
\]

**Proof.** We only take continuous domains situation into consideration and omit finite discrete domains situation, since the proof procedure is quite similar. The proof starts from the definition of \( R_{\Delta} \).

\[
R_{\Delta} = \int_X \mathcal{D}(x) \cdot 1[x \in D_{\alpha} \Delta D_h] \, dx
\]

\[
= \int_X (\mathcal{U}_X(x) + \mathcal{D}(x) - \mathcal{U}_X(x)) \cdot 1[x \in D_{\alpha} \Delta D_h] \, dx
\]

\[
= R_{\Delta,t} + \int_X (\mathcal{D}(x) - \mathcal{U}_X(x)) \cdot 1[x \in D_{\alpha} \Delta D_h] \, dx
\]

\[
\geq R_{\Delta,t} - \int_X \sup_x |\mathcal{D}(x) - \mathcal{U}_X(x)| \cdot 1[x \in D_{\alpha} \Delta D_h] \, dx
\]

\[
\geq R_{\Delta,t} - \sqrt{\frac{1}{2} D_{KL}(\mathcal{D}||\mathcal{U}_X)} \int_X 1[x \in D_{\alpha} \Delta D_h] \, dx
\]

\[
= R_{\Delta,t} - \#(D_{\alpha} \Delta D_h) \sqrt{\frac{1}{2} D_{KL}(\mathcal{D}||\mathcal{U}_X)}
\]

\[
\geq R_{\Delta,t} - \#X \sqrt{\frac{1}{2} D_{KL}(\mathcal{D}||\mathcal{U}_X)},
\]

where the second inequality is by the Pinsker’s inequality. \( \blacksquare \)

Denote \( \lambda \mathcal{T}_{h_t} + (1 - \lambda) \mathcal{U}_X \) as \( \mathcal{D}_t \), and \( R_{\Delta} + \#X \sqrt{\frac{1}{2} D_{KL}(\mathcal{D}_t||\mathcal{U}_X)} \) as \( \Psi^{R_{\Delta,t}}_{D_{KL}(\mathcal{D}_t||\mathcal{U}_X)} \). We now can bound \( |D_{h_t}| \) and \( |D_\epsilon \cap D_{h_t}| \) more explicitly.

**Lemma 7**

Given \( f \in \mathcal{F} \), \( \epsilon > 0 \) and any hypothesis \( h_t \in \mathcal{H} \), \( |D_{h_t}| \) is upper bounded by

\[
|D_{h_t}| \leq |D_{\alpha_\epsilon}| + \Psi^{R_{\Delta,t}}_{D_{KL}(\mathcal{D}_t||\mathcal{U}_X)},
\]
and $|D_e \cap D_{h_t}|$ is lower bounded by

$$|D_e \cap D_{h_t}| \geq |D_e| - 2\Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}.$$  

where $D_t = \lambda T_{h_t} + (1 - \lambda)U_X$, and $\Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)} = R_{D_t} + \#X\sqrt{\frac{1}{2}D_{KL}(D_t\||U_X)}$.

**Proof.** By Lemma 4 and Lemma 6, we have $|D_{h_t}| \leq |D_{\alpha_t}| + R_{D_t} + \#X\sqrt{\frac{1}{2}D_{KL}(D_t\||U_X)}$. By Lemma 5 and Lemma 6, we have $|D_e \cap D_{h_t}| \geq |D_e| - 2R_{D_t} - \#X\sqrt{\frac{1}{2}D_{KL}(D_t\||U_X)}$. ■

**Bounding $Pr_{h_t}$ Explicitly**

On the basis of Lemma 3 and Lemma 7, we are able to derive an explicit lower bound of $Pr_{h_t}$.

**Lemma 8**

Given $f \in F$, $\epsilon > 0$ and any hypothesis $h_t \in H$, $Pr_{h_t}$ is lower bounded by

$$Pr_{h_t} \geq \frac{|D_e| - 2\Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}}{|D_{\alpha_t}| + \Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}} - \#D_t\sqrt{\frac{1}{2}D_{KL}(T_{h_t}\||U_{D_{h_t}})},$$

where $T_{h_t}$ is the distribution defined on $D_{h_t}$ induced by $h_t$, $U_{D_{h_t}}$ is the uniform distribution over $D_{h_t}, D_t = \lambda T_{h_t} + (1 - \lambda)U_X$, and $\Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)} = R_{D_t} + \#X\sqrt{\frac{1}{2}D_{KL}(D_t\||U_X)}$.

**Proof.** By Lemma 3, we have $Pr_{h_t} \geq \frac{|D_e \cap D_{h_t}|}{|D_{\alpha_t}| + \Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}} - \#(D_e \cap D_{h_t})\sqrt{\frac{1}{2}D_{KL}(T_{h_t}\||U_{D_{h_t}})}$. Combining it with Lemma 7 results in that

$$Pr_{h_t} \geq \frac{|D_e| - 2\Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}}{|D_{\alpha_t}| + \Psi^{R_{D_t}}_{D_{KL}(D_t\||U_X)}} - \#D_t\sqrt{\frac{1}{2}D_{KL}(T_{h_t}\||U_{D_{h_t}})}.$$

■

**Proof of Lemma 2**

**Proof.** Since $D_t = \lambda U_{D_{h_t}} + (1 - \lambda)U_X$, we have $T_{h_t} = U_{D_{h_t}}$ and thus $D_{KL}(T_{h_t}\||U_{D_{h_t}}) = 0$. Now, combining the definition of $\overline{Pr}_h (\frac{1}{T} \sum_{t=1}^{T} Pr_{h_t})$ and Lemma 8 proves the theorem. ■

**Proof of Theorem 1**

**Theorem 1**

Given $f \in F$, $0 < \delta < 1$ and $\epsilon > 0$, if a classifier-based optimization algorithm has error-target $\theta$-dependence and $\gamma$-shrinking rate, its $(\epsilon, \delta)$-query complexity belongs to

$$O \left( \frac{1}{|D_e|} \left( (1 - \lambda) + \frac{\lambda}{\gamma T} \sum_{t=1}^{T} \frac{1 - Q \cdot R_{D_t} - \theta}{|D_{\alpha_t}|} \right)^{-1} \ln \frac{1}{\delta} \right),$$

where $Q = 1/(1 - \lambda)$.

To prove this theorem, our strategy is to refine the bound of $|D_e \cap D_{h_t}|$ under the error-target $\theta$-dependence condition and the bound of $|D_{h_t}|$ under the $\gamma$-shrinking rate condition, respectively.
Refining the Bounds of $|D_e \cap D_{h_t}|$ and $|D_{h_t}|$

**Lemma 9**

For the classifier-based optimization algorithms under the condition of error-target $\theta$-dependence,

$$|D_e \cap D_{h_t}| \geq |D_e| \cdot (1 - R_{\Delta X, t} - \theta)$$

holds for all $t$, where $R_{\Delta X, t}$ is the generalization error of $h_t$ under $\Delta X$ in iteration $t$.

**Proof.** Assume w.l.o.g. that $\epsilon \leq \alpha_t$ for all $t$, we have

$$|D_e \cap D_{h_t}| = |D_e| - |D_e \cap (D_{\alpha_t} \Delta D_{h_t})|$$

$$\geq |D_e| - |D_{\alpha_t} \Delta D_{h_t}| \cdot |\theta| |D_e|$$

$$= |D_e|(1 - |D_{\alpha_t} \Delta D_{h_t}| - \theta),$$

where the first equality is by $D_e \subseteq D_{\alpha_t}$, and the first inequality is by the condition of error-target $\theta$-dependence.

Let $R_{\Delta X, t}$ denote the generalization error of $h_t$ under $\Delta X$ in iteration $t$, it can be verified directly that $R_{\Delta X, t} = |D_{\alpha_t} \Delta D_{h_t}|$ under 0-1 loss. Thus, we have $|D_e \cap D_{h_t}| \geq |D_e|(1 - R_{\Delta X, t} - \theta)$. ■

In order to refine Lemma 9, i.e., lower bound $|D_e \cap D_{h_t}|$ using the generalization error of $h_t$ under the true sampling distribution $D_t = \lambda D_{h_t} + (1 - \lambda) D_X$ instead of $D_X$, we need Lemma 10 below.

It gives a relationship between $R_{\Delta X, t}$ and $R_{\Delta t}$, where $R_{\Delta t}$ is the generalization error of $h_t$ under $D_t$ in iteration $t$.

**Lemma 10**

For any $h_t \in H_t$, let $D_t = \lambda D_{h_t} + (1 - \lambda) D_X$, it holds for all $t$ that $R_{\Delta X, t} \leq R_{\Delta t}/(1 - \lambda)$, where $\lambda \in (0, 1)$.

**Proof.** We only consider continuous domains situation and omit finite discrete domains situation since the proof procedure is quite similar. Let $D_{\neq h_t}$ be the region where $h_t$ makes mistakes. Splitting $D_{\neq h_t}$ into $D_{\neq h_t}^+ = D_{\neq h_t} \cap D_{h_t}$ and $D_{\neq h_t}^t = D_{\neq h_t} - D_{\neq h_t}^+$, we can calculate the probability density $D_t(x) = \lambda \frac{1}{\#D_{h_t}} + (1 - \lambda) \frac{1}{\#X}$ for any $x \in D_{\neq h_t}$, and $D_t(x) = (1 - \lambda) \frac{1}{\#X}$ for any $x \in D_{\neq h_t}^t$. Thus,

$$R_{\Delta t} = \int_X D_t(x) \cdot 1[h_t \text{ makes mistake on } x] \, dx$$

$$= \int_{D_{\neq h_t}^t} D_t(x) \, dx = \int_{D_{\neq h_t}^+} D_t(x) \, dx + \int_{D_{\neq h_t}^t} D_t(x) \, dx$$

$$\geq \int_{D_{\neq h_t}^+} (1 - \lambda) \frac{1}{\#X} \, dx + \int_{D_{\neq h_t}^t} (1 - \lambda) \frac{1}{\#X} \, dx$$

$$= (1 - \lambda) R_{\Delta X, t},$$

which proves the lemma. ■

Let $Q = 1/(1 - \lambda)$. Combining Lemma 10 with Lemma 9, we can conclude that $|D_e \cap D_{h_t}| \geq |D_{h_t}| \cdot (1 - Q \cdot R_{\Delta t} - \theta)$. Meanwhile, the $\gamma$-shrinking rate condition admits $|D_{h_t}| \leq \gamma |D_{\alpha_t}|$ for all $t$ directly.

**Proof of Theorem 1**

**Proof.** By Lemma 3 and the assumption of $T_{h_t} = U_{D_{h_t}}$, we have $D_{KL}(T_{h_t} \| U_{D_{h_t}}) = 0$ and thus $\Pr_{h_t} \geq |D_e \cap D_{h_t}|/|D_{h_t}|$ for all $t$. Combining it with the refined bounds of $|D_e \cap D_{h_t}|$ and $|D_{h_t}|$ results in that $\Pr_{h_t} \geq \frac{(1 - Q \cdot R_{\Delta t} - \theta) \cdot |D_{\alpha_t}|}{|D_{\alpha_t}|}$, where $Q = 1/(1 - \lambda)$. Finally, by the definition of $\overline{\Pr}_{h_t}$ and Lemma 1 we prove the theorem. ■
Proof of Corollary 1

COROLLARY 1
In finite discrete domains $X = \{0, 1\}^n$, given $f \in F_L^{β_1, L_1, β_2, L_2}$, 0 < $δ < 1$ and $0 < \epsilon \leq L_1(\frac{2}{n})^{β_1}$, for a classifier-based optimization algorithm using a classification algorithm with convergence rate $\tilde{Θ}(\frac{1}{m})$, under the conditions that error-target dependence $\theta < 1$ and shrinking rate $γ > 0$, the $(ε, δ)$-query complexity of the classifier-based optimization algorithm belongs to $\text{poly}(\frac{1}{ε}, n, \frac{1}{β_1}, β_2, L_1, \ln L_1)$.

Proof. By the proof procedure of Theorem 1, letting $Q = 2$ (i.e., $λ = 1/2$), we have $\text{Pr}_h \geq \frac{1}{T} \sum_{t=1}^{T} \left( K_t \cdot |D_t| \right) / \left( \gamma \cdot |D_{α_t}| \right)$, where $K_t = 1 - 2R_{D_t} - θ$. Assume that $θ < 1$, since $K_t = (1 - 2R_{D_t} - θ)$ for all $t$, there must exist a constant $K > 0$ such that $K_t \geq K$ as long as $R_{D_t} < (1 - θ)/2$ for all $t$. Under the assumption of classifier-based optimization using the classification algorithms with convergence rate $\tilde{Θ}(\frac{1}{m})$, $R_{D_t} < (1 - θ)/2$ can be guaranteed if the sampled solution size $m$ in each iteration belongs to $\text{poly}(\frac{1}{ε}, n)$ [2]. Letting $K' = K/γ$, we therefore obtain that $\text{Pr}_h \geq \frac{1}{T} \sum_{t=1}^{T} \left( K_t \cdot |D_t| \right) / \left( γ \cdot |D_{α_t}| \right) = K' \sum_{t=1}^{T} |D_t| / |D_{α_t}|$.

Since $f \in F_L^{β_1, L_1, β_2, L_2}$, we know $L_2\|x - x^*\|_{β_2}^2 \leq f(x) - f(x^*) \leq L_1\|x - x^*\|_{β_1}^2$. Denote $\tilde{D}_t = \{ x \in X | \|x - x^*\|_{β_1}^2 \leq \frac{D_t}{T} \}$. It can be verified directly that $\tilde{D}_t \subseteq D_t$ and thus $|\tilde{D}_t| \leq |D_t|$. Let $α_t = f(x^*)$ and we assume that $α_t > 0$. $D_{α_t} = \{ x \in X | f(x) ≤ α_t \} = \{ x \in X | f(x) - f(x^*) ≤ α_t' \}$. Denote $\tilde{D}_{α_t} = \{ x \in X | \|x - x^*\|_{β_2}^2 ≤ \frac{α_t'}{T} \}$. Similarly, we have $D_{α_t} \subseteq \tilde{D}_{α_t}$ and thus $|D_{α_t}| \leq |\tilde{D}_{α_t}|$. For simplicity, we assume that $(\frac{n}{T})_{\frac{1}{β_1}}$ and $(\frac{n}{T})_{\frac{1}{β_2}}$ are both positive integers. By the definition of Hamming distance, we have

$$\#\tilde{D}_t = \sum_{i=0}^{(\frac{n}{T})_{\frac{1}{β_1}}} \binom{n}{i}, \quad \#\tilde{D}_{α_t} = \sum_{i=0}^{(\frac{n}{T})_{\frac{1}{β_2}}} \binom{n}{i}.$$ 

Let $H(p) = -p \log p - (1 - p) \log(1 - p)$ which is the binary entropy function of $p$, where $0 \leq p \leq 1$ and $H(p) = 0$ for $p = 0, 1$. Then, the following inequality [11] holds for all integers $0 \leq k \leq n$ with $p = k/n \leq 1/2$

$$\frac{1}{1 + \sqrt{8np(1 - p)}} \cdot 2^{nH(p)} \leq \sum_{i=0}^{n} \binom{n}{i} \leq 2^{nH(p)}.$$

Since $0 < \epsilon \leq L_1(\frac{2}{n})^{β_1}$, we have $(\frac{n}{T})_{\frac{1}{β_1}} \leq \frac{n}{T}$. Meanwhile, choosing $α'_t = (\frac{n}{T})_{\frac{1}{β_2}}$ for all $t$ can guarantee that $(\frac{n}{T})_{\frac{1}{β_2}} \leq \frac{n}{T}$ for all $t$ because $(\frac{n}{T})_{\frac{1}{β_2}} = 1 \leq (\frac{n}{T})_{\frac{1}{β_2}}$ for $n \geq 2$. If $n = 1$, we can still choose smaller $α'_t$ s.t. $(\frac{n}{T})_{\frac{1}{β_2}} \leq \frac{n}{T}$, and we omit the details since it is easy to verify. Combining the above statement with the inequality $\text{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T} \frac{|\tilde{D}_{α_t}|}{|D_{α_t}|}$, we have

$$\text{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T} \frac{|\tilde{D}_{α_t}|}{|D_{α_t}|} \geq \frac{K'}{T} \sum_{t=1}^{T} \binom{n}{(\frac{n}{T})_{\frac{1}{β_1}}} \cdot \frac{2^{nH(\frac{n}{T})_{\frac{1}{β_1}}}}{1 + \sqrt{8(\frac{n}{T})_{\frac{1}{β_1}}(1 - (\frac{n}{T})_{\frac{1}{β_1}}/n)}} \cdot \sum_{t=1}^{T} 2^{-nH(\frac{α'_t}{T})}.$$ 

Let the number of iterations $T$ to approach $(\frac{α'_t}{T})_{\frac{1}{β_2}} = (\frac{n}{T})_{\frac{1}{β_2}}$. Solving this equation results in that $T = \frac{β_2}{β_1} \log \frac{L_2}{ε} + 1 \in \text{poly}(\frac{1}{ε}, n, \frac{1}{β_1}, β_2, log L_1)$. For simplicity, we assume that $\frac{β_2}{β_1} \log \frac{L_2}{ε} + 1$ is a
positive integer and let the classifier-based optimization algorithms run \( T = \frac{\beta_T}{\beta_1} \log \frac{L_1^+}{\epsilon} + 1 \) number of iterations. Now, we can conclude that \( \mathbf{P}_{T_0} \geq \left( \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \right)^{-1} \).

Substituting \( \mathbf{P}_{T_0} \geq \left( \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \right)^{-1} \) into Lemma 1, we have \( (m + 1)T \in \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \cdot \ln \frac{1}{\delta}, \) with probability at least \( 1 - \delta \). Finally, combining the fact that \( R_{D_L} < (1 - \theta)/2 \) can be guaranteed by \( \text{poly}(\frac{1}{\varepsilon}, n) \) sampled solutions in each iteration and \( T \in \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \), the \((\varepsilon, \delta)\)-query complexity of the classifier-based optimization algorithms belongs to \( \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \cdot \ln \frac{1}{\delta}. \)


**Proof of Corollary 2**

**Corollary 2**

In compact continuous domains \( X \), given \( f \in F_{\beta_1, L_1, \beta_2, L_2} \), \( 0 < 1 < \epsilon > 0 \), for a classifier-based optimization algorithm using a classification algorithm with convergence rate \( \tilde{O}((\frac{1}{\epsilon})) \), under the conditions that error-target dependence \( \theta < 1 \) and shrinking rate \( \gamma > 0 \), the \((\varepsilon, \delta)\)-query complexity of the classifier-based optimization algorithm belongs to \( \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \cdot \ln \frac{1}{\delta}. \)

**Proof.** By the proof procedure of Theorem 1, letting \( Q = 2 \) (i.e., \( \lambda = 1/2 \)), we have \( \mathbf{P}_{T_0} \geq 1/2 \sum_{t=1}^{T} (K_t \cdot |D_t|) / (\gamma \cdot |D_{\alpha_t}|) \), where \( K_t = 1 - 2R_{D_L} - \theta \). Assume that \( \theta < 1 \), since \( K_t = 1 - 2R_{D_L} - \theta \) for all \( t \), there must exist a constant \( K > 0 \) such that \( K_t \geq K \) as long as \( R_{D_L} < (1 - \theta)/2 \) for all \( t \). Under the assumption of classifier-based optimization using the classification algorithms with convergence rate \( \tilde{O}(\frac{1}{\epsilon}) \), \( R_{D_L} < (1 - \theta)/2 \) can be guaranteed if the sampled solution size \( m \) in each iteration belongs to \( \text{poly}(\frac{1}{\varepsilon}, n) \) [2]. Letting \( K' = K/\gamma \), we therefore obtain that \( \mathbf{P}_{T_0} \geq 1/2 \sum_{t=1}^{T} (K \cdot |D_t|) / (\gamma \cdot |D_{\alpha_t}|) = K' \sum_{t=1}^{T} |D_t| / |D_{\alpha_t}| \).

Since \( f \in F_{\beta_1, L_1, \beta_2, L_2} \), we know \( L_2 \| x - x^* \|_{L_2}^{\beta_2} \leq f(x) - f(x^*) \leq L_1 \| x - x^* \|_{L_1}^{\beta_1} \). Denote \( \tilde{D}_t = \{ x \in X \mid \| x - x^* \|_{L_2}^{\beta_2} \leq \frac{T}{L_2} \} \). It can be verified directly that \( \tilde{D}_t \subseteq D_t \) and thus \( |\tilde{D}_t| \leq |D_t| \).

Let \( \alpha_t' = \alpha_t - f(x^*) \) and we assume that \( \alpha_t' > 0 \). \( D_{\alpha_t} = \{ x \in X \mid f(x) \leq \alpha_t \} = \{ x \in X \mid f(x) - f(x^*) \leq \alpha_t' \} \). Denote \( \tilde{D}_{\alpha_t} = \{ x \in X \mid \| x - x^* \|_{L_2}^{\beta_2} \leq \frac{\alpha_t'}{\epsilon^{\frac{\beta_2}{2}}} \} \). Similarly, we have \( D_{\alpha_t} \subseteq \tilde{D}_{\alpha_t} \) and thus \( |D_{\alpha_t}| \leq |\tilde{D}_{\alpha_t}| \). Note that \( \#\tilde{D}_t \) is the volume of \( \ell_2 \) ball of radius \( (\frac{T}{L_2})^{\frac{\beta_2}{2}} \) in \( \mathbb{R}^n \) which is proportional to \( (\frac{T}{L_2})^{\frac{n \beta_2}{2}} \), and \( \#\tilde{D}_{\alpha_t} \) is the volume of \( \ell_2 \) ball of radius \( (\frac{\alpha_t'}{\epsilon^{\frac{\beta_2}{2}}})^{\frac{1}{\beta_2}} \) in \( \mathbb{R}^n \) which is proportional to \( (\frac{\alpha_t'}{\epsilon^{\frac{\beta_2}{2}}})^{\frac{n \beta_2}{2}} \). Combing it with the inequality \( \mathbf{P}_{T_0} \geq K' \sum_{t=1}^{T} |D_t| / |D_{\alpha_t}| \), we have

\[
\mathbf{P}_{T_0} \geq \frac{K'}{T} \sum_{t=1}^{T} \frac{|\tilde{D}_t|}{|D_{\alpha_t}|} = \frac{K'}{T} \sum_{t=1}^{T} \frac{\#\tilde{D}_t}{\#D_{\alpha_t}} \\
= \frac{K'}{T} \sum_{t=1}^{T} \frac{(\epsilon/\gamma)^{\frac{n \beta_2}{2}}}{(\alpha_t'/\epsilon^{\frac{\beta_2}{2}})^{\frac{n}{\beta_2}}} \\
= \frac{K'}{T} \left( \frac{L_2^{\frac{\beta_2}{2}}}{L_1^{\frac{\beta_1}{2}} \epsilon^{\frac{\beta_1}{2}}} \right)^n \left( \frac{\alpha_t'}{\epsilon^{\frac{\beta_2}{2}}} \right)^{-\frac{n}{\beta_2}}.
\]

We choose \( \alpha_t' = \frac{1}{\beta_2} \), and use the number of iterations \( T \) to approach \( (\alpha_t')^{-\frac{n}{\beta_2}} = (\frac{L_2^{\frac{\beta_2}{2}}}{L_1^{\frac{\beta_1}{2}} \epsilon^{\frac{\beta_1}{2}}} \cdot (\frac{\beta_2}{\beta_1}))^{-n} \).

Solving this equation results in that \( T = \frac{\beta_2}{\beta_1} \log \frac{L_1^+}{\epsilon} - \log L_2 \in \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \).

For simplicity, we assume that \( \frac{\beta_2}{\beta_1} \log \frac{L_1^+}{\epsilon} - \log L_2 \) is a positive integer and let the classifier-based optimization algorithms run \( T = \frac{\beta_2}{\beta_1} \log \frac{L_1^+}{\epsilon} - \log L_2 \) number of iterations. Now, we can conclude that \( \mathbf{P}_{T_0} \geq \left( \text{poly}(\frac{1}{\varepsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_1}) \right)^{-1} \).
Substituting $\mathbf{Pr}_h \geq \left( \text{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{\epsilon_{\text{opt}}}) \right)^{-1}$ into Lemma 1, we have $(m + 1)T \in \text{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{\epsilon_{\text{opt}}}) \cdot \ln \frac{1}{\delta}$, with probability at least $1 - \delta$. Finally, combining the fact that $R_{T_D} < (1 - \theta)/2$ can be guaranteed with $\text{poly}(\frac{1}{\epsilon}, n)$ sampled solutions in each iteration and $T \in \text{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{\epsilon_{\text{opt}}})$, the $(\epsilon, \delta)$-query complexity of the classifier-based optimization algorithms belongs to $\text{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{\epsilon_{\text{opt}}}) \cdot \ln \frac{1}{\delta}$. \hfill \qed

**Proof of Corollary 3**

**Corollary 3**

In compact continuous domains $X$, given $f \in \mathcal{F}$ satisfying $\sum_{t=1}^{T}(\alpha_t')^{N_{\epsilon} - n} \in \Omega(\epsilon^{N_{\epsilon} - n})$, $0 < \delta < 1$ and $\epsilon > 0$, for a classifier-based optimization algorithm using the classification algorithms with convergence rate $\Theta(\frac{1}{m})$, under the conditions that error-dependence tolerance $\theta < 1$ and shrinking rate $\gamma > 0$, the $(\epsilon, \delta)$-query complexity of the classifier-based optimization algorithm belongs to $\text{poly}(\frac{1}{\epsilon}, n) \cdot \ln \frac{1}{\delta}$.

**Proof.** By the proof procedure of Theorem 1, letting $Q = 2$ (i.e., $\lambda = 1/2$), we have $\mathbf{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T}(K_t \cdot |D_t|)/(\gamma \cdot |D_{\alpha_t}|)$, where $K_t = 1 - 2R_{D_t} - \theta$. Assume that $\theta < 1$, since $K_t = 1 - 2R_{D_t} - \theta$ for all $t$, there must exist a constant $K > 0$ such that $K_t \geq K$ as long as $R_{D_t} < (1 - \theta)/2$ for all $t$. Under the assumption of classifier-based optimization using the classification algorithms with convergence rate $\Theta(\frac{1}{m})$, $R_{D_t} < (1 - \theta)/2$ can be guaranteed if the sampled solution size $m$ in each iteration belongs to $\text{poly}(\frac{1}{\epsilon}, n)$. Letting $K' = K/\gamma$, we therefore obtain that $\mathbf{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T}(K \cdot |D_t|)/(\gamma \cdot |D_{\alpha_t}|) = \frac{K'}{T} \sum_{t=1}^{T}|D_t|/|D_{\alpha_t}|$.

Recall that $D_t = \{x \in X | f(x) - f(x^*) \leq \epsilon\}$ for any $\epsilon > 0$. Let $\alpha_t' = \alpha_t - f(x^*)$ and we assume that $\alpha_t' > 0$, thus, $D_{\alpha_t} = \{x \in X | f(x) \leq \alpha_t\} = \{x \in X | f(x) - f(x^*) \leq \alpha_t'\}$. Let $V(D_t)$, $V(D_{\alpha_t})$ and $V(\eta\epsilon)$ denote the volume of $D_t$, $D_{\alpha_t}$ and $\ell_2$ ball of radius $\eta\epsilon$ in $\mathbb{R}^n$ respectively. By the definition of $N_p$ and $N_{\epsilon}$, we have

$$C_1 \epsilon^{-N_{\epsilon}} \cdot V(\eta\epsilon) \leq V(D_{\alpha_t}) = \#D_{\alpha_t} \leq C_2 \epsilon^{-N_{\epsilon}} \cdot V(\eta\epsilon).$$

Note that the volume of $\ell_2$ ball of radius $\epsilon$ in $\mathbb{R}^n$ is $\frac{\pi^{n/2}}{\Gamma(n/2 + 1)}(\eta\epsilon)^n$. Combing it with the inequality $\mathbf{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T}|D_t|/|D_{\alpha_t}|$, we have

$$\mathbf{Pr}_h \geq \frac{K'}{T} \sum_{t=1}^{T} \frac{|D_t|}{|D_{\alpha_t}|} = \frac{K'}{T} \sum_{t=1}^{T} \#D_{\alpha_t} \leq \frac{K'}{T} \sum_{t=1}^{T} C_2 \epsilon^{-N_{\epsilon}} \cdot V(\eta\epsilon)^n = \frac{K'}{T} \sum_{t=1}^{T} C_2 \epsilon^{-N_{\epsilon}} \cdot V(\eta\epsilon)^n = \frac{C_1 K'}{C_2} \sum_{t=1}^{T} \epsilon^{n-N_{\epsilon}} = \frac{C_1 K'}{C_2} \sum_{t=1}^{T} \epsilon^{n-N_{\epsilon}}. \ln \frac{1}{\delta}.$$

Let $T \in \text{poly}(\frac{1}{\epsilon}, n)$, if the problem $f \in \mathcal{F}$ satisfying $\sum_{t=1}^{T}(\alpha_t')^{N_{\epsilon} - n} \in \Omega(\epsilon^{N_{\epsilon} - n})$, we can conclude that $\mathbf{Pr}_h \geq (\text{poly}(\frac{1}{\epsilon}, n))^{-1}$.

Substituting $\mathbf{Pr}_h \geq (\text{poly}(\frac{1}{\epsilon}, n))^{-1}$ into Lemma 1, we have $(m + 1)T \in \text{poly}(\frac{1}{\epsilon}, n) \cdot \ln \frac{1}{\delta}$, with probability at least $1 - \delta$. Finally, combining the fact that $R_{T_D} < (1 - \theta)/2$ can be guaranteed with $\text{poly}(\frac{1}{\epsilon}, n)$ sampled solutions in each iteration and $T \in \text{poly}(\frac{1}{\epsilon}, n)$, the $(\epsilon, \delta)$-query complexity of the classifier-based optimization algorithms belongs to $\text{poly}(\frac{1}{\epsilon}, n) \cdot \ln \frac{1}{\delta}$. \hfill \qed

**References**