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 $f\colon X \to \mathbb{R}$ 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 $\#D = \int_{x\in X} \mathbb{I}[x \in D]\,\mathrm{d}x$ (or $\#D = \sum_{x\in X} \mathbb{I}[x \in D]$ for finite discrete domains), where $\mathbb{I}[\cdot]$ 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 Δ 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 $h\colon X \to \{-1,+1\}$. Let $\mathcal{H} \subseteq \{h\colon X \to \{-1,+1\}\}$ 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 $\mathcal{T}_h$ 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 $\texttt{poly}(\cdot)$ be the set of all polynomials w.r.t. the related variables and $\texttt{superpoly}(\cdot)$ be the set of all functions that grow faster than any function in $\texttt{poly}(\cdot)$ w.r.t. the related variables.

Proof of Lemma 1

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

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

where $\overline{\mathbf{Pr}}_h = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Pr}_{h_t} = \frac{1}{T}\sum_{t=1}^{T}\int_{D_\epsilon}\mathcal{U}_{D_{h_t}}(x)\,\mathrm{d}x$ (or $\overline{\mathbf{Pr}}_h = \frac{1}{T}\sum_{t=1}^{T}\sum_{x\in D_\epsilon}\mathcal{U}_{D_{h_t}}(x)$ for finite discrete domains) is the average success probability of sampling from the learned positive area of h_t, and $m_{\mathbf{Pr}_{h_t}}$ is the sample size required to realize the success probability $\mathbf{Pr}_{h_t}$.

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 $\tilde{x}$ s.t. $f(\tilde{x}) - f(x^*) > \epsilon$, and we denote it as $\mathbf{Pr}(f(\tilde{x}) - f(x^*) > \epsilon)$. Since $\tilde{x}$ is the best solution among all sampled solutions, $\mathbf{Pr}(f(\tilde{x}) - f(x^*) > \epsilon)$ 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 $1 - \mathbf{Pr}_u$, where $\mathbf{Pr}_u = |D_\epsilon| = \#D_\epsilon / \#X$ 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 $1 - \mathbf{Pr}_{h_t}$, where $\mathbf{Pr}_{h_t} = \int_{D_\epsilon} \mathcal{T}_{h_t}(x)\,\mathrm{d}x$ (or $\mathbf{Pr}_{h_t} = \sum_{x \in D_\epsilon} \mathcal{T}_{h_t}(x)$ 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

$$
\begin{aligned}
&\mathbf{Pr}(f(\tilde{x}) - f(x^*) > \epsilon) = (1 - \mathbf{Pr}_u)^m \cdot \prod_{t=1}^{T} \sum_{i=0}^{m} \binom{m}{i} (1-\lambda)^i \lambda^{m-i} (1 - \mathbf{Pr}_u)^i (1 - \mathbf{Pr}_{h_t})^{m-i} \\
&= (1 - \mathbf{Pr}_u)^m \prod_{t=1}^{T} \big((1-\lambda)(1 - \mathbf{Pr}_u) + \lambda(1 - \mathbf{Pr}_{h_t})\big)^m \\
&= (1 - \mathbf{Pr}_u)^m \prod_{t=1}^{T} \big(1 - (1-\lambda)\mathbf{Pr}_u - \lambda \mathbf{Pr}_{h_t}\big)^m \\
&\leq \exp\big(-\mathbf{Pr}_u \cdot m\big) \cdot \prod_{t=1}^{T} \exp\Big(-\big((1-\lambda)\mathbf{Pr}_u \cdot m + \lambda \mathbf{Pr}_{h_t} \cdot m\big)\Big) \\
&= \exp\Big(-\big(\mathbf{Pr}_u \cdot m + (1-\lambda)\sum_{t=1}^{T} \mathbf{Pr}_u \cdot m + \lambda \sum_{t=1}^{T} \mathbf{Pr}_{h_t} \cdot m\big)\Big) \\
&\leq \exp\Big(-\big((1-\lambda)\sum_{t=1}^{T} \mathbf{Pr}_u \cdot m + \lambda \sum_{t=1}^{T} \mathbf{Pr}_{h_t} \cdot m\big)\Big) \\
&= \exp\Big(-\big((1-\lambda)\mathbf{Pr}_u + \lambda \overline{\mathbf{Pr}}_h\big) \cdot mT\Big),
\end{aligned}
$$

where the first inequality is by $(1-x) \leq \exp(-x)$ for $x \in [0,1]$, and $\overline{\mathbf{Pr}}_h = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Pr}_{h_t}$.

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

$$\exp\Big(-\big((1-\lambda)\mathbf{Pr}_u + \lambda \overline{\mathbf{Pr}}_h\big) \cdot mT\Big) < \delta.$$

Therefore, we derive that $mT \in O\Big(\frac{1}{(1-\lambda)\mathbf{Pr}_u + \lambda \overline{\mathbf{Pr}}_h} \ln \frac{1}{\delta}\Big)$. At last, by $(m+1)T \leq 2mT \in O\Big(\frac{1}{(1-\lambda)\mathbf{Pr}_u + \lambda \overline{\mathbf{Pr}}_h} \ln \frac{1}{\delta}\Big)$ and $\mathbf{Pr}_u = |D_\epsilon| = \#D_\epsilon / \#X$, we prove the lemma. ■

Proof of Lemma 2

Let $R_{\mathcal{D}}$ denote the generalization error of $h \in \mathcal{H}$ with respect to the target function under distribution $\mathcal{D}$, 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 $\overline{\mathbf{Pr}}_h$ is lower bounded by

$$\overline{\mathbf{Pr}}_h \geq \frac{1}{T}\sum_{t=1}^{T} \Big(|D_\epsilon| - 2\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t \| \mathcal{U}_X)}\Big) \Big/ \Big(|D_{\alpha_t}| + \Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t \| \mathcal{U}_X)}\Big),$$

where $\mathcal{D}_t = \lambda\, \mathcal{U}_{D_{h_t}} + (1-\lambda)\mathcal{U}_X$ is the sampling distribution at iteration t, and $\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t \| \mathcal{U}_X)} = R_{\mathcal{D}_t} + \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}_t \| \mathcal{U}_X)}$.

To prove this lemma, our strategy is to first bound $\mathbf{Pr}_{h_t}$, which is the success probability of sampling from the distributions induced by the learned hypothesis at iteration t, and then bound $\overline{\mathbf{Pr}}_h$ by definition.

Bounding $\mathbf{Pr}_{h_t}$

In this section, we will bound $\mathbf{Pr}_{h_t}$ by two steps. A primary lower bound of $\mathbf{Pr}_{h_t}$ 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 $h_t \in \mathcal{H}$, $\mathbf{Pr}_{h_t}$ is lower bounded by

$$\mathbf{Pr}_{h_t} \geq \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} - \#(D_\epsilon \cap D_{h_t})\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}})},$$

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 $\mathbb{I}[\cdot]$ denote the indicator function, the proof starts from the definition of $\mathbf{Pr}_{h_t}$.

$$\begin{aligned}
\mathbf{Pr}_{h_t} &= \int_{D_{h_t}} \mathcal{T}_{h_t}(x) \cdot \mathbb{I}[x \in D_\epsilon]\,\mathrm{d}x = \int_{D_{h_t}} \big(\mathcal{T}_{h_t}(x) - \mathcal{U}_{D_{h_t}}(x) + \mathcal{U}_{D_{h_t}}(x)\big) \cdot \mathbb{I}[x \in D_\epsilon]\,\mathrm{d}x \\
&= \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} + \int_{D_{h_t}} \big(\mathcal{T}_{h_t}(x) - \mathcal{U}_{D_{h_t}}(x)\big) \cdot \mathbb{I}[x \in D_\epsilon]\,\mathrm{d}x \\
&\geq \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} - \int_{D_{h_t}} \sup_x |\mathcal{T}_{h_t}(x) - \mathcal{U}_{D_{h_t}}(x)| \cdot \mathbb{I}[x \in D_\epsilon]\,\mathrm{d}x \\
&\geq \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} - \sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}})} \int_{D_{h_t}} \mathbb{I}[x \in D_\epsilon]\,\mathrm{d}x \\
&= \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} - \#(D_\epsilon \cap D_{h_t})\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}})},
\end{aligned}$$

where $\mathcal{U}_{D_{h_t}}$ 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 $\mathbf{Pr}_{h_t}$, we need to investigate $|D_{h_t}|$ and $|D_\epsilon \cap D_{h_t}|$, 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

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

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

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

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

LEMMA 5
Given $f \in \mathcal{F}$, $\epsilon > 0$ and any hypothesis $h_t \in \mathcal{H}$, $|D_\epsilon \cap D_{h_t}|$ is lower bounded by

$$|D_\epsilon \cap D_{h_t}| \geq |D_\epsilon| - 2R_{\mathcal{U},t},$$

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

Proof. We w.l.o.g. assume that $\epsilon \le \alpha_t$ for all t. Let Δ denote the symmetric difference operator of two sets, by set operators, we have

$$
\begin{aligned}
&|D_\epsilon \cap D_{h_t}| = |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{h_t}| \\
&\ge |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{\alpha_t}| - |D_{\alpha_t} \Delta D_{h_t}| \\
&= |D_\epsilon \cup D_{h_t}| - |D_\epsilon \Delta D_{\alpha_t}| - R_{\mathcal{U},t} \\
&= |D_\epsilon \cup D_{h_t}| + |D_\epsilon| - |D_{\alpha_t}| - R_{\mathcal{U},t} \\
&\ge |D_{h_t}| + |D_\epsilon| - |D_{\alpha_t}| - R_{\mathcal{U},t},
\end{aligned}
$$

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

$$|D_\epsilon \cap D_{h_t}| \ge (|D_{h_t}| - |D_{\alpha_t}|) + |D_\epsilon| - R_{\mathcal{U},t} \ge |D_\epsilon| - 2R_{\mathcal{U},t}.$$

■

Bounding $|D_{h_t}|$ and $|D_\epsilon \cap D_h|$ More Explicitly

Lemma 4 and 5 show that $|D_{h_t}|$ and $|D_\epsilon \cap D_h|$ are bounded by the generalization error $R_{\mathcal{U}_X,t}$ 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 $R_{\mathcal{U}_X,t}$ and $R_{\mathcal{D}_t}$ in order to bound $|D_{h_t}|$ and $|D_\epsilon \cap D_h|$ more explicitly via $R_{\mathcal{D}_t}$.

LEMMA 6
The generalization error $R_{\mathcal{U}_X}$ of h under $\mathcal{U}_X$ and the generalization error $R_{\mathcal{D}}$ of h under any distribution $\mathcal{D}$ have the following relationship:

$$R_{\mathcal{U}_X} \le R_{\mathcal{D}} + \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}\|\mathcal{U}_X)}.$$

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 $R_{\mathcal{D}}$.

$$
\begin{aligned}
R_{\mathcal{D}} &= \int_X \mathcal{D}(x)\cdot \mathbb{I}[x \in D_\alpha \Delta D_h]\,\mathrm{d}x \\
&= \int_X \big(\mathcal{U}_X(x) + \mathcal{D}(x) - \mathcal{U}_X(x)\big)\cdot \mathbb{I}[x \in D_\alpha \Delta D_h]\,\mathrm{d}x \\
&= R_{\mathcal{U}_X} + \int_X \big(\mathcal{D}(x) - \mathcal{U}_X(x)\big)\cdot \mathbb{I}[x \in D_\alpha \Delta D_h]\,\mathrm{d}x \\
&\ge R_{\mathcal{U}_X} - \int_X \sup_x |\mathcal{D}(x) - \mathcal{U}_X(x)|\cdot \mathbb{I}[x \in D_\alpha \Delta D_h]\,\mathrm{d}x \\
&\ge R_{\mathcal{U}_X} - \sqrt{\frac{1}{2}D_{KL}(\mathcal{D}\|\mathcal{U}_X)}\int_X \mathbb{I}[x \in D_\alpha \Delta D_h]\,\mathrm{d}x \\
&= R_{\mathcal{U}_X} - \#(D_\alpha \Delta D_h)\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}\|\mathcal{U}_X)} \\
&\ge R_{\mathcal{U}_X} - \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}\|\mathcal{U}_X)},
\end{aligned}
$$

where the second inequality is by the Pinsker's inequality. ■

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

$$|D_{h_t}| \le |D_{\alpha_t}| + \Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)},$$

and $|D_\epsilon \cap D_{h_t}|$ *is lower bounded by*

$$|D_\epsilon \cap D_{h_t}| \geq |D_\epsilon| - 2\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)},$$

where $\mathcal{D}_t = \lambda \mathcal{T}_{h_t} + (1-\lambda)\mathcal{U}_X$, *and* $\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)} = R_{\mathcal{D}_t} + \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}$.

Proof. By Lemma 4 and Lemma 6, we have $|D_{h_t}| \leq |D_{\alpha_t}| + R_{\mathcal{D}_t} + \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}$. By Lemma 5 and Lemma 6, we have $|D_\epsilon \cap D_{h_t}| \geq |D_\epsilon| - 2R_{\mathcal{D}_t} - 2\#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}$. ■

Bounding $\mathbf{Pr}_{h_t}$ Explicitly

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

LEMMA 8
Given $f \in \mathcal{F}$, $\epsilon > 0$ *and any hypothesis* $h_t \in \mathcal{H}$, $\mathbf{Pr}_{h_t}$ *is lower bounded by*

$$\mathbf{Pr}_{h_t} \geq \frac{|D_\epsilon| - 2\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}}{|D_{\alpha_t}| + \Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}} - \#D_\epsilon\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t}\|\mathcal{U}_{D_{h_t}})},$$

where $\mathcal{T}_{h_t}$ *is the distribution defined on* D_{h_t} *induced by* h_t, $\mathcal{U}_{D_{h_t}}$ *is the uniform distribution over* D_{h_t}, $\mathcal{D}_t = \lambda \mathcal{T}_{h_t} + (1-\lambda)\mathcal{U}_X$, *and* $\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)} = R_{\mathcal{D}_t} + \#X\sqrt{\frac{1}{2}D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}$.

Proof. By Lemma 3, we have $\mathbf{Pr}_{h_t} \geq \frac{|D_\epsilon \cap D_{h_t}|}{|D_{h_t}|} - \#(D_\epsilon \cap D_{h_t})\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t}\|\mathcal{U}_{D_{h_t}})}$. Combining it with Lemma 7 results in that

$$\begin{aligned}\mathbf{Pr}_{h_t} &\geq \frac{|D_\epsilon| - 2\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}}{|D_{\alpha_t}| + \Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}} - \#(D_\epsilon \cap D_{h_t})\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t}\|\mathcal{U}_{D_{h_t}})} \\ &\geq \frac{|D_\epsilon| - 2\Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}}{|D_{\alpha_t}| + \Psi^{R_{\mathcal{D}_t}}_{D_{KL}(\mathcal{D}_t\|\mathcal{U}_X)}} - \#D_\epsilon\sqrt{\frac{1}{2}D_{KL}(\mathcal{T}_{h_t}\|\mathcal{U}_{D_{h_t}})}.\end{aligned}$$

■

Proof of Lemma 2

Proof. Since $\mathcal{D}_t = \lambda\,\mathcal{U}_{D_{h_t}} + (1-\lambda)\mathcal{U}_X$, we have $\mathcal{T}_{h_t} = \mathcal{U}_{D_{h_t}}$ and thus $D_{KL}(\mathcal{T}_{h_t}\|\mathcal{U}_{D_{h_t}}) = 0$. Now, combining the definition of $\overline{\mathbf{Pr}}_h$ ($= \frac{1}{T}\sum_{t=1}^{T}\mathbf{Pr}_{h_t}$) and Lemma 8 proves the theorem. ■

Proof of Theorem 1

THEOREM 1
Given $f \in \mathcal{F}$, $0 < \delta < 1$ *and* $\epsilon > 0$, *if a classifier-based optimization algorithm has error-target* θ*-dependence and* γ*-shrinking rate, its* (ϵ,δ)*-query complexity belongs to*

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

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

To prove this theorem, our strategy is to refine the bound of $|D_\epsilon \cap D_{h_t}|$ under the *error-target* θ*-dependence* condition and the bound of $|D_{h_t}|$ under the γ*-shrinking rate* condition, respectively.

Refining the Bounds of $|D_\epsilon \cap D_{h_t}|$ and $|D_{h_t}|$

LEMMA 9
For the classifier-based optimization algorithms under the condition of error-target θ-dependence,

$$|D_\epsilon \cap D_{h_t}| \geq |D_\epsilon| \cdot (1 - R_{\mathcal{U}_X,t} - \theta)$$

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

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

$$\begin{aligned}
&|D_\epsilon \cap D_{h_t}| = |D_\epsilon| - |D_\epsilon \cap (D_{\alpha_t} \Delta D_{h_t})| \\
&\geq |D_\epsilon| - |D_\epsilon| \cdot |D_{\alpha_t} \Delta D_{h_t}| - \theta |D_\epsilon| \\
&= |D_\epsilon|(1 - |D_{\alpha_t} \Delta D_{h_t}| - \theta),
\end{aligned}$$

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

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

In order to refine Lemma 9, i.e., lower bound $|D_\epsilon \cap D_{h_t}|$ using the generalization error of h_t under the true sampling distribution $\mathcal{D}_t = \lambda \mathcal{U}_{D_{h_t}} + (1-\lambda)\mathcal{U}_X$ instead of $\mathcal{U}_X$, we need Lemma 10 below. It gives a relationship between $R_{\mathcal{U}_X,t}$ and $R_{\mathcal{D}_t}$, where $R_{\mathcal{D}_t}$ is the generalization error of h_t under $\mathcal{D}_t$ in iteration t.

LEMMA 10
For any $h_t \in \mathcal{H}$, let $\mathcal{D}_t = \lambda \mathcal{U}_{D_{h_t}} + (1-\lambda)\mathcal{U}_X$, it holds for all t that $R_{\mathcal{U}_X,t} \leq R_{\mathcal{D}_t}/(1-\lambda)$, 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 $D_{\neq,t}$ be the region where h_t makes mistakes. Splitting $D_{\neq,t}$ into $D^+_{\neq,t} = D_{\neq,t} \cap D_{h_t}$ and $D^-_{\neq,t} = D_{\neq,t} - D^+_{\neq,t}$, we can calculate the probability density $\mathcal{D}_t(x) = \lambda \frac{1}{\#D_{h_t}} + (1-\lambda)\frac{\#D_{h_t}}{\#X}\frac{1}{\#D_{h_t}} = \lambda \frac{1}{\#D_{h_t}} + (1-\lambda)\frac{1}{\#X}$ for any $x \in D^+_{\neq,t}$, and $\mathcal{D}_t(x) = (1-\lambda)\frac{\#(X-D_{h_t})}{\#X}\frac{1}{\#(X-D_{h_t})} = (1-\lambda)\frac{1}{\#X}$ for any $x \in D^-_{\neq,t}$. Thus,

$$\begin{aligned}
R_{\mathcal{D}_t} &= \int_X \mathcal{D}_t(x) \cdot \mathbb{I}[h_t \text{ makes mistake on } x]\,\mathrm{d}x \\
&= \int_{D_{\neq,t}} \mathcal{D}_t(x)\,\mathrm{d}x = \int_{D^+_{\neq,t}} \mathcal{D}_t(x)\,\mathrm{d}x + \int_{D^-_{\neq,t}} \mathcal{D}_t(x)\,\mathrm{d}x \\
&\geq \int_{D^+_{\neq,t}} (1-\lambda)\frac{1}{\#X}\,\mathrm{d}x + \int_{D^-_{\neq,t}} (1-\lambda)\frac{1}{\#X}\,\mathrm{d}x \\
&= (1-\lambda)R_{\mathcal{U}_X,t},
\end{aligned}$$

which proves the lemma. ■

Let $Q = 1/(1-\lambda)$. Combining Lemma 10 with Lemma 9, we can conclude that $|D_\epsilon \cap D_{h_t}| \geq |D_\epsilon| \cdot (1 - Q \cdot R_{\mathcal{D}_t} - \theta)$. 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 $\mathcal{T}_{h_t} = \mathcal{U}_{D_{h_t}}$, we have $D_{KL}(\mathcal{T}_{h_t} \| \mathcal{U}_{D_{h_t}}) = 0$ and thus $\mathbf{Pr}_{h_t} \geq |D_\epsilon \cap D_{h_t}|/|D_{h_t}|$ for all t. Combining it with the refined bounds of $|D_\epsilon \cap D_{h_t}|$ and $|D_{h_t}|$ results in that $\mathbf{Pr}_{h_t} \geq \frac{(1-Q\cdot R_{\mathcal{D}_t} - \theta)\cdot|D_\epsilon|}{\gamma \cdot |D_{\alpha_t}|}$, where $Q = 1/(1-\lambda)$. Finally, by the definition of $\overline{\mathbf{Pr}}_h$ and Lemma 1 we prove the theorem. ■

Proof of Corollary 1

COROLLARY 1
In finite discrete domains $X = \{0,1\}^n$, given $f \in \mathcal{F}_L^{\beta_1,L_1,\beta_2,L_2}$, $0 < \delta < 1$ and $0 < \epsilon \leq L_1(\frac{n}{2})^{\beta_1}$, for a classifier-based optimization algorithm using a classification algorithm with convergence rate $\widetilde{\Theta}(\frac{1}{m})$, under the conditions that error-target dependence $\theta < 1$ and shrinking rate $\gamma > 0$, the (ϵ,δ)-query complexity of the classifier-based optimization algorithm belongs to $\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln\frac{1}{L_2}) \cdot \ln\frac{1}{\delta}$.

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

Since $f \in \mathcal{F}_L^{\beta_1,L_1,\beta_2,L_2}$, we know $L_2\|x - x^*\|_H^{\beta_2} \leq f(x) - f(x^*) \leq L_1\|x - x^*\|_H^{\beta_1}$. Denote $\widetilde{D}_\epsilon = \{x \in X \mid \|x - x^*\|_H^{\beta_1} \leq \frac{\epsilon}{L_1}\}$. It can be verified directly that $\widetilde{D}_\epsilon \subseteq D_\epsilon$ and thus $|\widetilde{D}_\epsilon| \leq |D_\epsilon|$. 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 $\widetilde{D}_{\alpha_t} = \{x \in X \mid \|x - x^*\|_H^{\beta_2} \leq \frac{\alpha'_t}{L_2}\}$. Similarly, we have $D_{\alpha_t} \subseteq \widetilde{D}_{\alpha_t}$ and thus $|D_{\alpha_t}| \leq |\widetilde{D}_{\alpha_t}|$. For simplicity, we assume that $(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}$ and $(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}}$ are both positive integers. By the definition of Hamming distance, we have

$$\#\widetilde{D}_\epsilon = \sum_{i=0}^{(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}}\binom{n}{i}, \quad \#\widetilde{D}_{\alpha_t} = \sum_{i=0}^{(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}}}\binom{n}{i}.$$

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 [1] holds for all integers $0 \leq k \leq n$ with $p = k/n \leq 1/2$

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

Since $0 < \epsilon \leq L_1(\frac{n}{2})^{\beta_1}$, we have $(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}} \leq \frac{n}{2}$. Meanwhile, choosing $\alpha'_t = \frac{2L_2}{2^t}$ for all t can guarantee that $(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}} \leq \frac{n}{2}$ for all t because $(\frac{\alpha'_1}{L_2}) = 1 \leq (\frac{n}{2})^{\beta_2}$ for $n \geq 2$. If $n = 1$, we can still choose smaller α'_t s.t. $(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}} \leq \frac{n}{2}$, and we omit the details since it is easy to verify. Combing the above statement with the inequality $\overline{\mathbf{Pr}}_h \geq \frac{K'}{T}\sum_{t=1}^{T}|D_\epsilon|/|D_{\alpha_t}|$, we have

$$\begin{aligned}
\overline{\mathbf{Pr}}_h &\geq \frac{K'}{T}\sum_{t=1}^{T}\frac{|\widetilde{D}_\epsilon|}{|\widetilde{D}_{\alpha_t}|} = \frac{K'}{T}\sum_{t=1}^{T}\frac{\#\widetilde{D}_\epsilon}{\#\widetilde{D}_{\alpha_t}} \\
&= \frac{K'}{T}\sum_{t=1}^{T}\frac{\sum_{i=0}^{(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}}\binom{n}{i}}{\sum_{i=0}^{(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}}}\binom{n}{i}} \\
&\geq \frac{K'}{T} \cdot \frac{2^{nH\left((\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}\right)}}{1 + \sqrt{8(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}\left(1 - (\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}/n\right)}}\sum_{t=1}^{T}2^{-nH\left((\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}}\right)}.
\end{aligned}$$

Let the number of iterations T to approach $(\frac{\alpha'_T}{L_2})^{\frac{1}{\beta_2}} = (\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}$. Solving this equation results in that $T = \frac{\beta_2}{\beta_1}\log\frac{L_1}{\epsilon} + 1 \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1)$. For simplicity, we assume that $\frac{\beta_2}{\beta_1}\log\frac{L_1}{\epsilon} + 1$ is a positive integer and let the classifier-based optimization algorithms run $T = \frac{\beta_2}{\beta_1}\log\frac{L_1}{\epsilon} + 1$ number of iterations. Now, we can conclude that $\overline{\mathbf{Pr}}_h \geq \left(\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log\frac{1}{L_2})\right)^{-1}$.

Substituting $\overline{\mathbf{Pr}}_h \geq \left(\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log\frac{1}{L_2})\right)^{-1}$ into Lemma 1, we have $(m+1)T \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln\frac{1}{L_2}) \cdot \ln\frac{1}{\delta}$, with probability at least $1-\delta$. Finally, combining the fact that $R_{\mathcal{D}_t} < (1-\theta)/2$ can be guaranteed with $\texttt{poly}(\frac{1}{\epsilon}, n)$ sampled solutions in each iteration and $T \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1)$, the ($\epsilon$,$\delta$)-query complexity of the classifier-based optimization algorithms belongs to $\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln\frac{1}{L_2}) \cdot \ln\frac{1}{\delta}$. ■

Proof of Corollary 2

COROLLARY 2
In compact continuous domains X, given $f \in \mathcal{F}_L^{\beta_1, L_1, \beta_2, L_2}$, $0 < \delta < 1$ and $\epsilon > 0$, for a classifier-based optimization algorithm using a classification algorithm with convergence rate $\widetilde{\Theta}(\frac{1}{m})$, under the conditions that error-target dependence $\theta < 1$ and shrinking rate $\gamma > 0$, the (ϵ,δ)-query complexity of the classifier-based optimization algorithm belongs to $\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln\frac{1}{L_2}) \cdot \ln\frac{1}{\delta}$.

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

Since $f \in \mathcal{F}_L^{\beta_1, L_1, \beta_2, L_2}$, we know $L_2\|x - x^*\|_2^{\beta_2} \leq f(x) - f(x^*) \leq L_1\|x - x^*\|_2^{\beta_1}$. Denote $\widetilde{D}_\epsilon = \{x \in X \mid \|x - x^*\|_2^{\beta_1} \leq \frac{\epsilon}{L_1}\}$. It can be verified directly that $\widetilde{D}_\epsilon \subseteq D_\epsilon$ and thus $|\widetilde{D}_\epsilon| \leq |D_\epsilon|$. 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 $\widetilde{D}_{\alpha_t} = \{x \in X \mid \|x - x^*\|_2^{\beta_2} \leq \frac{\alpha'_t}{L_2}\}$. Similarly, we have $D_{\alpha_t} \subseteq \widetilde{D}_{\alpha_t}$ and thus $|D_{\alpha_t}| \leq |\widetilde{D}_{\alpha_t}|$. Note that $\#\widetilde{D}_\epsilon$ is the volume of ℓ_2 ball of radius $(\frac{\epsilon}{L_1})^{\frac{1}{\beta_1}}$ in $\mathbb{R}^n$ which is proportional to $(\frac{\epsilon}{L_1})^{\frac{n}{\beta_1}}$, and $\#\widetilde{D}_{\alpha_t}$ is the volume of ℓ_2 ball of radius $(\frac{\alpha'_t}{L_2})^{\frac{1}{\beta_2}}$ in $\mathbb{R}^n$ which is proportional to $(\frac{\alpha'_t}{L_2})^{\frac{n}{\beta_2}}$. Combing it with the inequality $\overline{\mathbf{Pr}}_h \geq \frac{K'}{T}\sum_{t=1}^{T}|D_\epsilon|/|D_{\alpha_t}|$, we have

$$
\begin{aligned}
\overline{\mathbf{Pr}}_h &\geq \frac{K'}{T}\sum_{t=1}^{T}\frac{|\widetilde{D}_\epsilon|}{|\widetilde{D}_{\alpha_t}|} = \frac{K'}{T}\sum_{t=1}^{T}\frac{\#\widetilde{D}_\epsilon}{\#\widetilde{D}_{\alpha_t}} \\
&= \frac{K'}{T}\sum_{t=1}^{T}\frac{(\epsilon/L_1)^{\frac{n}{\beta_1}}}{(\alpha'_t/L_2)^{\frac{n}{\beta_2}}} \\
&= \frac{K'}{T}\cdot\left(\frac{L_2^{\frac{1}{\beta_2}}}{L_1^{\frac{1}{\beta_1}}}\epsilon^{\frac{1}{\beta_1}}\right)^n \sum_{t=1}^{T}(\alpha'_t)^{-\frac{n}{\beta_2}}.
\end{aligned}
$$

We choose $\alpha'_t = \frac{1}{2^t}$, and use the number of iterations T to approach $(\alpha'_T)^{-\frac{n}{\beta_2}} = (L_2^{\frac{1}{\beta_2}}\epsilon^{\frac{1}{\beta_1}}/L_1^{\frac{1}{\beta_1}})^{-n}$. Solving this equation results in that $T = \frac{\beta_2}{\beta_1}\log\frac{L_1}{\epsilon} - \log L_2 \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log\frac{1}{L_2})$. 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 $\overline{\mathbf{Pr}}_h \geq \left(\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log\frac{1}{L_2})\right)^{-1}$.

Substituting $\overline{\mathbf{Pr}}_h \geq \left(\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \log L_1, \log \frac{1}{L_2})\right)^{-1}$ into Lemma 1, we have $(m+1)T \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{L_2}) \cdot \ln \frac{1}{\delta}$, with probability at least $1-\delta$. Finally, combining the fact that $R_{\mathcal{D}_t} < (1-\theta)/2$ can be guaranteed with $\texttt{poly}(\frac{1}{\epsilon}, n)$ sampled solutions in each iteration and $T \in \texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{L_2})$, the ($\epsilon$,$\delta$)-query complexity of the classifier-based optimization algorithms belongs to $\texttt{poly}(\frac{1}{\epsilon}, n, \frac{1}{\beta_1}, \beta_2, \ln L_1, \ln \frac{1}{L_2}) \cdot \ln \frac{1}{\delta}$. ■

Proof of Corollary 3

COROLLARY 3
In compact continuous domains X, given $f \in \mathcal{F}$ satisfying $\sum_{t=1}^{T} (\alpha'_t)^{\mathcal{N}_c - n} \in \Omega\left(\epsilon^{\mathcal{N}_p - n}\right)$, $0 < \delta < 1$ and $\epsilon > 0$, for a classifier-based optimization algorithm using the classification algorithms with convergence rate $\widetilde{\Theta}(\frac{1}{m})$, under the conditions that error-target dependence $\theta < 1$ and shrinking rate $\gamma > 0$, the (ϵ,δ)-query complexity of the classifier-based optimization algorithm belongs to $\texttt{poly}(\frac{1}{\epsilon}, n) \cdot \ln \frac{1}{\delta}$.

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

Recall that $D_\epsilon = \{x \in X \mid f(x) - f(x^*) \leq \epsilon\}$ for any $\epsilon > 0$. Let $\alpha'_t = \alpha_t - f(x^*)$ and we assume that $\alpha'_t > 0$, thus, $D_{\alpha_t} = \{x \in X \mid f(x) \leq \alpha_t\} = \{x \in X \mid f(x) - f(x^*) \leq \alpha'_t\}$. Let $V(D_\epsilon)$, $V(D_{\alpha_t})$ and $V(\eta\epsilon)$ denote the volume of D_ϵ, D_{α_t} and ℓ_2 ball of radius $\eta\epsilon$ in $\mathbb{R}^n$ respectively. By the definition of $\mathcal{N}_p$ and $\mathcal{N}_c$, we have

$$C_1\epsilon^{-\mathcal{N}_p} \cdot V(\eta\epsilon) \leq V(D_\epsilon) = \#D_\epsilon \leq C_2\epsilon^{-\mathcal{N}_c} \cdot V(\eta\epsilon),$$
$$C_1(\alpha'_t)^{-\mathcal{N}_p} \cdot V(\eta\alpha'_t) \leq V(D_{\alpha_t}) = \#D_{\alpha_t} \leq C_2(\alpha'_t)^{-\mathcal{N}_c} \cdot V(\eta\alpha'_t).$$

Note that the volume of ℓ_2 ball of radius $\eta\epsilon$ in $\mathbb{R}^n$ is $\frac{\pi^{n/2}}{\Gamma(n/2+1)}(\eta\epsilon)^n$. Combing it with the inequality $\overline{\mathbf{Pr}}_h \geq \frac{K'}{T}\sum_{t=1}^{T}|D_\epsilon|/|D_{\alpha_t}|$, we have

$$\begin{aligned}
\overline{\mathbf{Pr}}_h &\geq \frac{K'}{T}\sum_{t=1}^{T}\frac{|D_\epsilon|}{|D_{\alpha_t}|} = \frac{K'}{T}\sum_{t=1}^{T}\frac{\#D_\epsilon}{\#D_{\alpha_t}} \\
&\geq \frac{K'}{T}\sum_{t=1}^{T}\frac{C_1\epsilon^{-\mathcal{N}_p} \cdot V(\eta\epsilon)}{C_2(\alpha'_t)^{-\mathcal{N}_c} \cdot V(\eta\alpha'_t)} = \frac{K'}{T}\sum_{t=1}^{T}\frac{C_1\epsilon^{-\mathcal{N}_p} \cdot (\eta\epsilon)^n}{C_2(\alpha'_t)^{-\mathcal{N}_c} \cdot (\eta\alpha'_t)^n} \\
&= \frac{C_1K'}{C_2T}\sum_{t=1}^{T}\frac{\epsilon^{n-\mathcal{N}_p}}{(\alpha'_t)^{n-\mathcal{N}_c}} = \frac{C_1K' \cdot \epsilon^{n-\mathcal{N}_p}}{C_2T}\sum_{t=1}^{T}(\alpha'_t)^{\mathcal{N}_c - n}.
\end{aligned}$$

Let $T \in \texttt{poly}(\frac{1}{\epsilon}, n)$, if the problem $f \in \mathcal{F}$ satisfying $\sum_{t=1}^{T} (\alpha'_t)^{\mathcal{N}_c - n} \in \Omega\left(\epsilon^{\mathcal{N}_p - n}\right)$, we can conclude that $\overline{\mathbf{Pr}}_h \geq \left(\texttt{poly}(\frac{1}{\epsilon}, n)\right)^{-1}$.

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

References

[1] R. B. Ash. *Information Theory*. Dover Publications Inc., New York, 1990.

[2] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1994.