A Related Definitions

This section presents background information on the Matern covariance function, differential entropy, and information gain.

A.1 Matern Covariance Function

The Matern covariance function, widely used in BO, is defined as

$$k_{\text{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}d}{l} \right)^{\nu} B_{\nu} \left(\frac{\sqrt{2\nu}d}{l} \right), \tag{23}$$

where $l > 0$, $d = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ represents the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, $\nu > 0$ denotes the smoothness parameter, $\Gamma(\cdot)$ represents the gamma function, and $B_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind. Varying ν determines the smoothness of samples drawn from a Gaussian process with this covariance function. Smaller values of ν correspond to rougher samples. Additionally, these samples are $\lceil \nu \rceil - 1$ times continuously differentiable (Williams & Rasmussen, 2006). Figure 3 illustrates samples drawn from a Gaussian process with this covariance function using different values of ν.

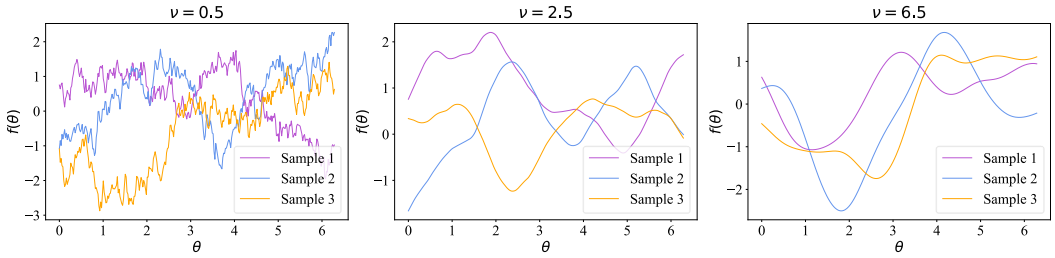

Figure 3: Samples drawn from a Gaussian process with the Matern covariance function $k_{\text{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ using smoothness parameters ν from $\nu = 0.5$ to $\nu = 6.5$.

A.2 Differential Entropy

Let X be a random variable with a probability density function q whose support is a set $\mathcal{X}$. The differential entropy $H(X)$ is defined as

$$\mathrm{H}[X] = \mathbb{E}[-\log(q(X))] = \int_{\mathcal{X}} q(x) \log q(x) dx. \tag{24}$$

Specifically, the differential entropy of a multivariate Gaussian random variable X_{Gaussian} with distribution $N(\boldsymbol{\mu}, \boldsymbol{K})$ is expressed as

$$\mathrm{H}[X_{\text{Gaussian}}] = \frac{1}{2} \log(\det(2\pi e \boldsymbol{K})), \tag{25}$$

where $\boldsymbol{\mu}$ denotes the mean vector and $\boldsymbol{K}$ represents the covariance matrix.

A.3 Information Gain

Let $\mathcal{S}_T = \{(\boldsymbol{\theta}_1, y(\boldsymbol{\theta}_1)), \cdots, (\boldsymbol{\theta}_T, y(\boldsymbol{\theta}_T))\}$ be T accumulated observations about the function $f(\boldsymbol{\theta})$, where $y(\boldsymbol{\theta}_t)$ denotes the estimation of $f(\boldsymbol{\theta}_t)$ for $t \in [T]$. The informativeness of $\mathcal{S}_T$ regarding $f(\boldsymbol{\theta})$ is quantified by the information gain g_T, which is the mutual information (Shannon, 1948) between $\boldsymbol{y}_T = [y(\boldsymbol{\theta}_1) \cdots y(\boldsymbol{\theta}_T)]^{\mathsf{T}}$ and $\boldsymbol{f}_T = [f(\boldsymbol{\theta}_1) \cdots f(\boldsymbol{\theta}_T)]^{\mathsf{T}}$. Specifically,

$$g_T = \mathrm{H}[\boldsymbol{y}_T] - \mathrm{H}[\boldsymbol{y}_T | \boldsymbol{f}_T], \tag{26}$$

where $\mathrm{H}[\boldsymbol{y}_T]$ represents the information entropy of $\boldsymbol{y}_T$ and $\mathrm{H}[\boldsymbol{y}_T | \boldsymbol{f}_T]$ denotes the conditional information entropy of $\boldsymbol{y}_T$ given $\boldsymbol{f}_T$.

B PROOF OF LEMMA 4.3

In this section, we present a complete proof of Lemma 4.3 through a sequence of lemmas. We initially establish the following result regarding the partial derivative $\partial_j f(\boldsymbol{\theta})$ of the noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ for any $j \in [2p]$ and any $\boldsymbol{\theta} \in \mathcal{D}$.

Lemma B.1. *Assuming that Assumption 4.2 holds, let $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the noise-free PS-QNN objective function. Given a failure probability $\delta \in (0, 1)$, the partial derivative $\partial_j f(\boldsymbol{\theta})$ satisfies*

$$\forall j \in [2p], \forall \boldsymbol{\theta} \in \mathcal{D},\ |\partial_j f(\boldsymbol{\theta})| \leq \sqrt{\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta} \tag{27}$$

with a success probability of at least $\geq 1 - \delta$, where $\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ is the variance of $\partial_a f(\boldsymbol{\theta})$ with index $a = \arg\max_{j \in [2p]} (\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|)$.

Proof. Fix $a \in [2p]$, by Chebyshev's Inequality, we have

$$\Pr\{\forall \boldsymbol{\theta} \in \mathcal{D}, \forall s > 0,\ |\partial_a f(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]| \leq s\} \geq 1 - \mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/s^2, \tag{28}$$

where $\mathbb{E}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ and $\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ are the expectation value and the variance of $\partial_a f(\boldsymbol{\theta})$. Assuming that Assumption 4.2 holds, we demonstrate that $\mathbb{E}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})] = 0$. The detailed proof can be found in Ref. Cerezo et al. (2021b). This implies

$$\Pr\{\forall \boldsymbol{\theta} \in \mathcal{D}, \forall s > 0,\ |\partial_a f(\boldsymbol{\theta})| \leq s\} \geq 1 - \mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/s^2. \tag{29}$$

By choosing $a = \arg\max_{j \in [2p]} (\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|)$, we have

$$\Pr\left\{\forall s > 0,\ \sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_a f(\boldsymbol{\theta})| \leq s\right\} \geq 1 - \mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/s^2. \tag{30}$$

The use of the index a and the notation $\sup(\cdot)$ immediately implies

$$\Pr\{\forall j \in [2p], \forall \boldsymbol{\theta} \in \mathcal{D}, \forall s > 0,\ |\partial_j f(\boldsymbol{\theta})| \leq s\} \geq 1 - \mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/s^2. \tag{31}$$

Let the failure probability $\delta = \mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/s^2 \in (0, 1)$, we have

$$\Pr\left\{\forall j \in [2p], \forall \boldsymbol{\theta} \in \mathcal{D},\ |\partial_j f(\boldsymbol{\theta})| \leq \sqrt{\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta}\right\} \geq 1 - \delta. \tag{32}$$

□

Lemma B.2. *Given a noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$, we have*

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D},\ |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')| \leq \max_{j \in [2p]} \left(\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|\right) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1, \tag{33}$$

where $\partial_j f(\boldsymbol{\theta})$ is the partial derivative of $f(\boldsymbol{\theta})$ for $j \in [2p]$.

Proof. Let $\boldsymbol{\theta}$ be represented as $[\theta_1, \cdots, \theta_{2p}]^\mathsf{T}$. For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D}$, we have

$$\begin{aligned} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') =& f(\theta_1, \cdots, \theta_{2p}) - f(\theta'_1, \theta_2, \cdots, \theta_{2p}) + \cdots + \\ & f(\theta'_1, \cdots, \theta'_{j-1}, \theta_j, \cdots, \theta_{2p}) - f(\theta'_1, \cdots, \theta'_j, \theta_{j+1}, \cdots, \theta_{2p}) + \cdots + \\ & f(\theta'_1, \cdots, \theta'_{2p-1}, \theta_{2p}) - f(\theta'_1, \cdots, \theta'_{2p}). \end{aligned} \tag{34}$$

By Triangle Inequality, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D}$, we have

$$\begin{aligned} |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')| \leq& |f(\theta_1, \cdots, \theta_{2p}) - f(\theta'_1, \theta_2, \cdots, \theta_{2p})| + \cdots + \\ & \left|f(\theta'_1, \cdots, \theta'_{j-1}, \theta_j, \cdots, \theta_{2p}) - f(\theta'_1, \cdots, \theta'_j, \theta_{j+1}, \cdots, \theta_{2p})\right| + \cdots + \\ & \left|f(\theta'_1, \cdots, \theta'_{2p-1}, \theta_{2p}) - f(\theta'_1, \cdots, \theta'_{2p})\right|. \end{aligned} \tag{35}$$

For any $j \in [2p]$, the partial derivative with respect to the problem-oriented Hamiltonian H_1

$$\partial_j f(\boldsymbol{\theta}) = i\langle\varphi_0|U_-^\dagger[H_1, U_+^\dagger H_1 U_+]U_-|\varphi_0\rangle \tag{36}$$

and the partial derivative with respect to the mixing Hamiltonian H_2

$$\partial_j f(\boldsymbol{\theta}) = i\langle\varphi_0|U_-^\dagger[H_2, U_+^\dagger H_1 U_+]U_-|\varphi_0\rangle \tag{37}$$

exist and are continuous on $\mathcal{D} = [0, 2\pi]^{2p}$, where U_- is the left slice circuit and U_+ is the right slice circuit of the variational parameter θ_j in the noise-free PS-QNN $U(\boldsymbol{\theta})$, and $|\varphi_0\rangle$ is the initial state. Fix $[\theta_1', \cdots, \theta_{j-1}', \theta_{j+1}, \cdots, \theta_{2p}]^\mathsf{T} \in [0, 2\pi]^{2p-1}$, $f(\boldsymbol{\theta})$ can be seen as an uni-variable function in θ_j. By Lagrange's Mean Value Theorem (Sohrab, 2003), for any $\theta_j, \theta_j' \in [0, 2\pi]$ and for any $[\theta_1', \cdots, \theta_{j-1}', \theta_{j+1}, \cdots, \theta_{2p}]^\mathsf{T} \in [0, 2\pi]^{2p-1}$ denoted as $\hat{\boldsymbol{\theta}} \in \hat{\mathcal{D}}$, we have

$$\left|f(\theta_1', \cdots, \theta_{j-1}', \theta_j, \cdots, \theta_{2p}) - f(\theta_1', \cdots, \theta_j', \theta_{j+1}, \cdots, \theta_{2p})\right| \leq L_{j,\hat{\boldsymbol{\theta}}} \left|\theta_j - \theta_j'\right|, \tag{38}$$

where $L_{j,\hat{\boldsymbol{\theta}}} = \sup_{\theta_j \in [0,2\pi]} |\partial_j f(\boldsymbol{\theta})|$. In light of this, for any $\theta_j, \theta_j' \in [0, 2\pi]$ and for any $\hat{\boldsymbol{\theta}} \in \hat{\mathcal{D}}$, we have

$$\left|f(\theta_1', \cdots, \theta_{j-1}', \theta_j, \cdots, \theta_{2p}) - f(\theta_1', \cdots, \theta_j', \theta_{j+1}, \cdots, \theta_{2p})\right| \leq L_j \left|\theta_j - \theta_j'\right|, \tag{39}$$

where $L_j = \sup_{\hat{\boldsymbol{\theta}} \in \hat{\mathcal{D}}} L_{j,\hat{\boldsymbol{\theta}}}$. Therefore, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D}$, we have

$$\begin{aligned}
|f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')| &\leq L_1 |\theta_1 - \theta_1'| + \cdots + L_{2p} \left|\theta_{2p} - \theta_{2p}'\right| && (40)\\
&\leq \left(\max_{j \in [2p]} L_j\right) \sum_{j=1}^{2p} \left|\theta_j - \theta_j'\right| && (41)\\
&= \max_{j \in [2p]} L_j \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 && (42)\\
&= \max_{j \in [2p]} \left(\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|\right) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1. && (43)
\end{aligned}$$

□

Given Lemma B.1 and Lemma B.2, we come to Lemma 4.3 straightforwardly.

Proof of Lemma 4.3. By Lemma B.1, we pick $\delta \in (0, 1)$ and have

$$\Pr\left\{\max_{j \in [2p]} \left(\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|\right) \leq \sqrt{\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta}\right\} \geq 1 - \delta, \tag{44}$$

where $\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ is the variance of the partial derivative $\partial_a f(\boldsymbol{\theta})$ with index $a = \arg\max_{j \in [2p]}(\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|)$. Substituting this into Lemma B.2, the statement holds. □

C Proof of Theorem 4.5

Theorem C.1 (Formal)**.** *Given a constant threshold ϵ, a failure probability $\delta \in (0, 1)$ and an n-qubit noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ induced by the network $U(\boldsymbol{\theta})$ that satisfies Assumption 4.2, run BO for $T = \mathrm{poly}(n^{1/\epsilon^2})$ steps, where the scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as*

$$\eta_t = 2\log(2\pi^2 t^2/3\delta) + 4p\log(8\pi p t^2 \sqrt{\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta}). \tag{45}$$

If the parameter dimension

$$p \leq \tilde{\mathcal{O}}\left(\sqrt{\log n}\right), \tag{46}$$

then the optimization error r_T satisfies $r_T \leq \epsilon$ with a success probability of at least $1 - \delta$. Here, $\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ is the variance of the partial derivative $\partial_a f(\boldsymbol{\theta})$ with index $a = \arg\max_{j \in [2p]}(\sup_{\boldsymbol{\theta} \in \mathcal{D}} |\partial_j f(\boldsymbol{\theta})|)$.

C.1 Outline of the Proof Procedure

Our objective is to determine the effective parameter dimension p of the noise-free PS-QNN $U(\boldsymbol{\theta})$ such that the optimization error $r_T = f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_T^+)$ after $T = \text{poly}(n)$ steps of executing BO can be upper bounded by a constant threshold ϵ. Here, $\boldsymbol{\theta}^*$ represents the global maximum point and $\boldsymbol{\theta}_T^+$ denotes the approximation of the maximum point in the previous T steps. We investigate this question through the perspective of the Bayesian approach, which considers the corresponding noise-free PS-QNN objective function $f(\boldsymbol{\theta})$ as a sample drawn from a Gaussian process with the Matern covariance function $k_{\text{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ (Eq. 23). We first establish that r_T is upper bounded by $\frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t))$, where $\boldsymbol{\theta}_t$ represents the next point selected in each step t. It is evident that the condition $\frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t)) \leq \epsilon$ is sufficient to deduce the result $r_T \leq \epsilon$. Hence, by ensuring that the upper bound on $\frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t))$ is no greater than ϵ, we can determine the effective p that guarantees $r_T \leq \epsilon$. Subsequently, we utilize the continuity property of the noise-free PS-QNN objective function $f(\boldsymbol{\theta})$ (Lemma 4.3) to establish an upper bound on $\frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t))$.

The complete proof of Theorem 4.5 is supported by a series of lemmas (Lemma C.2-Lemma C.8). We will introduce how these lemmas are employed in our proof. For convenience, we initially present explanations of several notions that commonly occur in the following sections. Specifically, $\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]$ denotes the variance of the partial derivative $\partial_a f(\boldsymbol{\theta})$ with index $a = \arg\max_{j\in[2p]}(\sup_{\boldsymbol{\theta}\in\mathcal{D}}|\partial_j f(\boldsymbol{\theta})|)$. Additionally, $\mu_{t-1}(\boldsymbol{\theta})$ represents the posterior mean function of $f(\boldsymbol{\theta})$ and $\sigma_{t-1}(\boldsymbol{\theta})$ denotes the posterior standard deviation of $f(\boldsymbol{\theta})$ based on the accumulated observations $\mathcal{S}_{t-1}$ from the previous $t-1$ steps.

To facilitate the analysis in the continuous domain $\mathcal{D} = [0, 2\pi]^{2p}$, we discretize $\mathcal{D}$ into a finite grid $\mathcal{D}_t$ in each step t, as it has been employed in Ref. Srinivas et al. (2012). Specifically, the size of $\mathcal{D}_t$ is determined by the degree of discretization τ_t, such that $|\mathcal{D}_t| = (\tau_t)^{2p}$. In the subsequent discussion, we use $[\boldsymbol{\theta}^*]_t$ to denote the closest point in $\mathcal{D}_t$ to $\boldsymbol{\theta}^*$. Next, we will evaluate upper bounds on $f(\boldsymbol{\theta}^*) - f([\boldsymbol{\theta}^*]_t)$ (the first term) and $f([\boldsymbol{\theta}^*]_t)$ (the second term) to obtain an upper bound on $f(\boldsymbol{\theta}^*)$. Regarding the first term, according to Lemma C.2, if $\tau_t = 8\pi p t^2\sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta}$, then $f(\boldsymbol{\theta}^*) - f([\boldsymbol{\theta}^*]_t)$ can be upper bounded by $1/t^2$ with a success probability of at least $1-\delta/4$. Considering that $\boldsymbol{\theta}_t$ is selected by maximizing the acquisition function $\text{UCB}_t(\boldsymbol{\theta})$ over $\mathcal{D}$, according to Lemma C.3, $\text{UCB}_t(\boldsymbol{\theta}_t) = \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)$ can be used to upper bound $f([\boldsymbol{\theta}^*]_t)$ with a success probability of at least $1-\delta/4$. Here, a predefined scaling parameter $\eta_t = 2\log\left(2\pi^2 t^2 |\mathcal{D}_t| / 3\delta\right)$ is used. Taking the two upper bounds mentioned above into account, Lemma C.4 demonstrates that

$$f(\boldsymbol{\theta}^*) = (f(\boldsymbol{\theta}^*) - f([\boldsymbol{\theta}^*]_t)) + f([\boldsymbol{\theta}^*]_t) \leq 1/t^2 + \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)$$

with a success probability of at least $1-\delta/2$. Furthermore, we establish that $f(\boldsymbol{\theta}_t)$ is lower bounded by $\mu_{t-1}(\boldsymbol{\theta}_t) - \sqrt{\eta_t'}\sigma_{t-1}(\boldsymbol{\theta}_t)$ with a success probability of at least $1-\delta/2$ using Lemma C.5, where $\eta_t' = 2\log(\pi^2 t^2/3\delta)$. Since $\eta_t \geq \eta_t'$, we can also use $\mu_{t-1}(\boldsymbol{\theta}_t) - \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)$ as a lower bound for $f(\boldsymbol{\theta}_t)$. Afterward, Lemma C.6 establishes that

$$f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t) \leq 1/t^2 + 2\sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)$$

with a success probability of at least $1-\delta$. Then, Lemma C.7 establishes a connection between the sum of posterior variances $\sum_{t=1}^{T}\sigma_{t-1}^2(\boldsymbol{\theta}_t)$ and the information gain g_T (Eq. 26). As $f(\boldsymbol{\theta})$ is considered as a sample drawn from a Gaussian process with $k_{\text{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, we can bound $\sum_{t=1}^{T}\sigma_{t-1}^2(\boldsymbol{\theta}_t)$ by the upper bound $\mathcal{O}(T^{\frac{p}{\nu+p}}\log^{\frac{\nu}{\nu+p}}(T))$ on the maximal g_T for $k_{\text{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in Ref. Vakili et al. (2021). By applying Cauchy-Schwarz Inequality and considering the non-decreasing property of η_t as t increases, we can substitute the form of η_T to obtain the result stated in Lemma C.8

$$r_T \leq \mathcal{O}\left(\sqrt{p\log\left(pT^2(\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})])^{1/2}\right)(\log T/T)^{\frac{\nu}{\nu+p}}}\right)$$

with a success probability of at least $1-\delta$. Finally, we obtain the effective p by solving for this upper bound is no greater than a constant threshold ϵ with $T = \text{poly}(n^{1/\epsilon^2})$.

C.2 Proof Details

In this section, we provide a comprehensive introduction to the corresponding lemmas.

Lemma C.2. *Assuming that Assumption 4.2 holds, let $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the n-qubit noise-free PS-QNN objective function. Given a failure probability $\delta \in (0, 1)$ and a finite grid $\mathcal{D}_t$ of size $|\mathcal{D}_t| = (\tau_t)^{2p}$ with the degree of discretization $\tau_t = 4\pi p t^2 \sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta}$ in each step t, run BO for $T = \mathrm{poly}(n)$ steps. The following relationship*

$$\forall t \in [T], \forall \boldsymbol{\theta} \in \mathcal{D},\ |f(\boldsymbol{\theta}) - f([\boldsymbol{\theta}]_t)| \leq 1/t^2 \tag{47}$$

holds with a success probability of at least $1 - \delta$, where $[\boldsymbol{\theta}]_t$ represents the closest point in $\mathcal{D}_t$ to $\boldsymbol{\theta}$.

Proof. By choosing a finite grid $\mathcal{D}_t$ of size $(\tau_t)^{2p}$ in each step t, for any $\boldsymbol{\theta} \in \mathcal{D}$ we have $\|\boldsymbol{\theta} - [\boldsymbol{\theta}]_t\|_1 \leq 4\pi p/\tau_t$. Given Lemma 4.3, we have

$$\Pr\left\{\forall t \in [T], \forall \boldsymbol{\theta} \in \mathcal{D},\ |f(\boldsymbol{\theta}) - f([\boldsymbol{\theta}]_t)| \leq 4\pi p \sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta}/\tau_t\right\} \geq 1 - \delta, \tag{48}$$

where the failure probability $\delta \in (0, 1)$. Since $\tau_t = 4\pi p t^2 \sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta}$, then

$$\Pr\left\{\forall t \in [T], \forall \boldsymbol{\theta} \in \mathcal{D},\ |f(\boldsymbol{\theta}) - f([\boldsymbol{\theta}]_t)| \leq 1/t^2\right\} \geq 1 - \delta. \tag{49}$$

Furthermore, we consider $\mathbb{V}[\partial_a f(\boldsymbol{\theta})]$ to be $1/\mathrm{poly}(n)$, as shown in Ref. Park & Killoran (2024). Additionally, we assume that parameter dimension p is at most $\mathrm{poly}(n)$. In order to guarantee the degree of discretization τ_t of at least 1, we enforce a constraint that the number of steps $T = \mathrm{poly}(n)$. This constraint is consistent with the scenario we are exploring. □

Lemma C.3. *Given a failure probability $\delta \in (0, 1)$, an n-qubit noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ and a finite grid $\mathcal{D}_t \subset \mathcal{D}$ of size $|\mathcal{D}_t|$ in each step t, run BO for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta_t = 2\log(\pi^2 t^2 |\mathcal{D}_t|/6\delta)$. The following relationship*

$$\forall t \in [T], \forall \boldsymbol{\theta} \in \mathcal{D}_t,\ f(\boldsymbol{\theta}) \in \mathcal{C}_t(\boldsymbol{\theta}) \tag{50}$$

holds with a success probability of at least $1 - \delta$, where $\mathcal{C}_t(\boldsymbol{\theta})$ represents a confidence interval $[\mu_{t-1}(\boldsymbol{\theta}) - \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}),\ \mu_{t-1}(\boldsymbol{\theta}) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta})]$.

Proof. Fix $t \in [T]$ and $\boldsymbol{\theta} \in \mathcal{D}_t$. Conditioned on accumulated observations $\mathcal{S}_{t-1}$ from the previous $t - 1$ steps, the posterior distribution $f(\boldsymbol{\theta}) \sim N(\mu_{t-1}(\boldsymbol{\theta}), \sigma^2_{t-1}(\boldsymbol{\theta}))$. Now, if $b \sim N(0, 1)$, then

$$\begin{aligned}
\Pr\{b > w\} &= \exp(-w^2/2)(2\pi)^{-1/2} \exp\left(-(b - w)^2/2 - w(b - w)\right) && (51)\\
&\leq \exp(-w^2/2) \Pr\{b > 0\} && (52)\\
&= \frac{1}{2}\exp(-w^2/2) && (53)
\end{aligned}$$

for $w > 0$, since $\exp(-w(b - w)) \leq 1$ for $b \geq w$. Using $b = (f(\boldsymbol{\theta}) - \mu_{t-1}(\boldsymbol{\theta}))/\sigma_{t-1}(\boldsymbol{\theta})$ and $w = \sqrt{\eta_t}$, we have

$$\Pr\{f(\boldsymbol{\theta}) \notin \mathcal{C}_t(\boldsymbol{\theta})\} \leq \exp(-\eta_t/2). \tag{54}$$

Applying the union bound for $\boldsymbol{\theta} \in \mathcal{D}_t$, we have

$$\Pr\{\forall \boldsymbol{\theta} \in \mathcal{D}_t,\ f(\boldsymbol{\theta}) \in \mathcal{C}_t(\boldsymbol{\theta})\} \geq 1 - |\mathcal{D}_t| \exp(-\eta_t/2). \tag{55}$$

Given that $|\mathcal{D}_t| \exp(-\eta_t/2) = \delta/q_t$, where $\sum_{t \geq 1}(1/q_t) = 1$, $q_t > 0$, by applying the union bound for $t \in \mathbb{N}$, the statement holds. For example, we can use $q_t = \pi^2 t^2/6$. □

Lemma C.4. *Assuming that Assumption 4.2 holds, let $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the n-qubit noise-free PS-QNN objective function. Given a failure probability $\delta \in (0, 1)$, run BO for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta_t = 2\log(\pi^2 t^2/3\delta) + 4p\log(4\pi p t^2 \sqrt{2\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta})$. The following relationship*

$$\forall t \in [T],\ f(\boldsymbol{\theta}^*) \leq \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t) + 1/t^2 \tag{56}$$

holds with a success probability of at least $1 - \delta$, where $\boldsymbol{\theta}^*$ denotes the global maximum point and $\boldsymbol{\theta}_t$ represents the next point selected in each step t.

Proof. Using the failure probability $\delta/2$ in Lemma C.2, for the global maximum point $\boldsymbol{\theta}^*$, we have

$$\Pr\{\forall t \in [T],\ f(\boldsymbol{\theta}^*) - f([\boldsymbol{\theta}^*]_t) \leq 1/t^2\} \geq 1 - \delta/2, \tag{57}$$

where $[\boldsymbol{\theta}^*]_t$ denotes the closest point in $\mathcal{D}_t$ to $\boldsymbol{\theta}^*$. Here, a finite grid $\mathcal{D}_t$ of size $|\mathcal{D}_t| = (\tau_t)^{2p}$ with $\tau_t = 4\pi p t^2 \sqrt{2\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta}$. Then, applying Lemma C.3 with the failure probability $\delta/2$, for $[\boldsymbol{\theta}^*]_t$, we have

$$\Pr\{\forall t \in [T],\ f([\boldsymbol{\theta}^*]_t) \leq \mu_{t-1}([\boldsymbol{\theta}^*]_t) + \sqrt{\eta_t}\sigma_{t-1}([\boldsymbol{\theta}^*]_t)\} \geq 1 - \delta/2, \tag{58}$$

where $\eta_t = 2\log(\pi^2 t^2 |\mathcal{D}_t| /3\delta)$. As the next point $\boldsymbol{\theta}_t$ is selected by maximizing $\mathrm{UCB}_t(\boldsymbol{\theta})$ in each step t, we have $\mathrm{UCB}_t([\boldsymbol{\theta}^*]_t) \leq \mathrm{UCB}_t(\boldsymbol{\theta}_t)$. Then, we have

$$\Pr\{\forall t \in [T],\ f([\boldsymbol{\theta}^*]_t) \leq \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)\} \geq 1 - \delta/2. \tag{59}$$

Taking Eq. 57 and Eq. 59 together, the statement holds since $(1 - \delta/2)^2 > 1 - \delta$. □

Lemma C.5. *Given a failure probability $\delta \in (0, 1)$ and an n-qubit noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$, run BO for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η'_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta'_t = 2\log(\pi^2 t^2/6\delta)$. The following relationship*

$$\forall t \in [T],\ f(\boldsymbol{\theta}_t) \in \mathcal{C}_t(\boldsymbol{\theta}_t) \tag{60}$$

holds with a success probability of at least $1 - \delta$, where $\boldsymbol{\theta}_t$ represents the next point selected in each step t and $\mathcal{C}_t(\boldsymbol{\theta}_t)$ denotes the confidence interval $[\mu_{t-1}(\boldsymbol{\theta}_t) - \sqrt{\eta'_t}\sigma_{t-1}(\boldsymbol{\theta}_t),\ \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta'_t}\sigma_{t-1}(\boldsymbol{\theta}_t)]$.

Proof. Fix $t \in [T]$. Conditioned on $\mathcal{S}_{t-1}$ from the previous $t-1$ steps, for the next point $\boldsymbol{\theta}_t$ selected in each step t, the posterior distribution $f(\boldsymbol{\theta}_t) \sim N(\mu_{t-1}(\boldsymbol{\theta}_t), \sigma^2_{t-1}(\boldsymbol{\theta}_t))$. Now, if $b \sim N(0, 1)$, then $\Pr\{b > w\} \leq \frac{1}{2}\exp(-w^2/2)$ for $w > 0$. Using $b = (f(\boldsymbol{\theta}_t) - \mu_{t-1}(\boldsymbol{\theta}_t))/\sigma_{t-1}(\boldsymbol{\theta}_t)$ and $w = \sqrt{\eta'_t}$, we have

$$\Pr\{f(\boldsymbol{\theta}_t) \notin \mathcal{C}_t(\boldsymbol{\theta}_t)\} \leq \exp(-\eta'_t/2). \tag{61}$$

Given that $\exp(-\eta'_t/2) = \delta/q_t$, where $\sum_{t\geq 1}(1/q_t) = 1$, $q_t > 0$, by applying the union bound for $t \in \mathbb{N}$, the statement holds. For example, we can use $q_t = \pi^2 t^2/6$. □

Lemma C.6. *Assuming that Assumption 4.2 holds, let $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the n-qubit noise-free PS-QNN objective function. Given a failure probability $\delta \in (0, 1)$, run BO for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta_t = 2\log(2\pi^2 t^2/3\delta) + 4p\log(8\pi p t^2\sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta})$. The following relationship*

$$\forall t \in [T],\ f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t) \leq 2\sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t) + 1/t^2 \tag{62}$$

holds with a success probability of at least $1 - \delta$, where $\boldsymbol{\theta}^*$ denotes the global maximum point and $\boldsymbol{\theta}_t$ represents the next point selected in each step t.

Proof. Using the failure probability $\delta/2$ in Lemma C.4, for the global maximum point $\boldsymbol{\theta}^*$, we have

$$\Pr\{\forall t \in [T],\ f(\boldsymbol{\theta}^*) \leq \mu_{t-1}(\boldsymbol{\theta}_t) + \sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t) + 1/t^2\} \geq 1 - \delta/2 \tag{63}$$

with $\eta_t = 2\log(2\pi^2 t^2/3\delta) + 4p\log(8\pi p t^2\sqrt{\mathbb{V}[\partial_a f(\boldsymbol{\theta})]/\delta})$ in each step t. Then, using the failure probability $\delta/2$ in Lemma C.5, for the next point $\boldsymbol{\theta}_t$ selected in each step t, we have

$$\Pr\{\forall t \in [T],\ f(\boldsymbol{\theta}_t) \geq \mu_{t-1}(\boldsymbol{\theta}_t) - \sqrt{\eta'_t}\sigma_{t-1}(\boldsymbol{\theta}_t)\} \geq 1 - \delta/2 \tag{64}$$

with $\eta'_t = 2\log(\pi^2 t^2/3\delta)$ in each step t. As the aforementioned η_t is greater than η'_t used here, choosing η_t here is also valid. Taking Eq. 63 and Eq. 64 together, the proof is completed. □

Lemma C.7. *Given an n-qubit noise-free PS-QNN objective function $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$, run BO for $T = \mathrm{poly}(n)$ steps. Let $\mathcal{S}_T = \{(\boldsymbol{\theta}_1, y(\boldsymbol{\theta}_1)), \cdots, (\boldsymbol{\theta}_T, y(\boldsymbol{\theta}_T))\}$ be the accumulated observations from the previous T steps, where the estimation $y(\boldsymbol{\theta}_t) = f(\boldsymbol{\theta}_t) + \xi_t^{\mathrm{noise}}$ in each step t. Here, $\xi_t^{\mathrm{noise}} \sim N(0, 1/4M)$ is independent and identically distributed Gaussian noise with M representing the fixed number of measurements. The information gain g_T (Eq. 26) can be expressed as*

$$g_T = \frac{1}{2}\sum_{t=1}^{T}\log(1 + 4M\sigma^2_{t-1}(\boldsymbol{\theta}_t)). \tag{65}$$

Proof. Let $\boldsymbol{y}_{t-1} = [y(\boldsymbol{\theta}_1) \cdots y(\boldsymbol{\theta}_{t-1})]^\mathsf{T}$ and $\boldsymbol{f}_{t-1} = [f(\boldsymbol{\theta}_1) \cdots f(\boldsymbol{\theta}_{t-1})]^\mathsf{T}$ for $t \in [T+1]$. Plugging in the differential entropy of a multivariate Gaussian random variable (Eq. 25), we have $\mathrm{H}[y(\boldsymbol{\theta}_t)|\boldsymbol{y}_{t-1}] = 1/2 \log(2\pi e(1/4M + \sigma^2_{t-1}(\boldsymbol{\theta}_t)))$ for $t \in [T]$ and $\mathrm{H}[\boldsymbol{y}_T|\boldsymbol{f}_T] = \frac{T}{2}\log(\pi e/2M)$. Using the fact that $\mathrm{H}[\boldsymbol{y}_t] = \mathrm{H}[\boldsymbol{y}_{t-1}] + \mathrm{H}[y(\boldsymbol{\theta}_t)|\boldsymbol{y}_{t-1}]$, we have

$$\mathrm{H}[\boldsymbol{y}_T] = \mathrm{H}[\boldsymbol{y}_0] + \mathrm{H}[y(\boldsymbol{\theta}_1)|\boldsymbol{y}_0] + \mathrm{H}[y(\boldsymbol{\theta}_2)|\boldsymbol{y}_1] + \cdots + \mathrm{H}[y(\boldsymbol{\theta}_T)|\boldsymbol{y}_{T-1}] \tag{66}$$

$$= \frac{1}{2}\sum_{t=1}^{T} \log(2\pi e(1/4M + \sigma^2_{t-1}(\boldsymbol{\theta}_t))). \tag{67}$$

Recalling the definition of g_T (Eq. 26), the statement holds. □

Lemma C.8. *Assuming that Assumption 4.2 holds, let $f(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the n-qubit noise-free PS-QNN objective function. Given a failure probability $\delta \in (0, 1)$, run BO with the Matern prior covariance function $k_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ (Eq. 23) for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta_t = 2\log(2\pi^2 t^2/3\delta) + 4p\log(8\pi p t^2 \sqrt{\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})]/\delta})$. The optimization error r_T satisfies*

$$r_T \leq \mathcal{O}\left(\sqrt{p \log\left(pT^2(\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})])^{1/2}\right)(\log T/T)^{\frac{\nu}{\nu+p}}}\right) \tag{68}$$

with a success probability of at least $1 - \delta$.

Proof. Noted that η_t in Lemma C.6 is non-decreasing. Since $0 \leq 4M\sigma^2_{t-1}(\boldsymbol{\theta}_t) \leq 4Mk_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) \leq 4M$, denoted as $4M\sigma^2_{t-1}(\boldsymbol{\theta}_t) \in [0, 4M]$, we have $4M\sigma^2_{t-1}(\boldsymbol{\theta}_t) \leq (4M/\log(1+4M))\log(1+4M\sigma^2_{t-1}(\boldsymbol{\theta}_t))$. Moreover, Lemma C.7 links the sum of the posterior variances $\sum_{t=1}^{T}\sigma^2_{t-1}(\boldsymbol{\theta}_t)$ to the information gain g_T. By Cauchy-Schwarz Inequality, we have

$$\left(\sum_{t=1}^{T} 2\sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t)\right)^2 \leq \sum_{t=1}^{T} 4\eta_t \sum_{t=1}^{T}\sigma^2_{t-1}(\boldsymbol{\theta}_t) \tag{69}$$

$$\leq \frac{T\eta_T}{M}\sum_{t=1}^{T}(4M\sigma^2_{t-1}(\boldsymbol{\theta}_t)) \tag{70}$$

$$\leq \frac{4T\eta_T}{\log(1+4M)}\sum_{t=1}^{T}\log(1+4M\sigma^2_{t-1}(\boldsymbol{\theta}_t)) \tag{71}$$

$$= c_0 T\eta_T g_T, \tag{72}$$

where the parameter $c_0 = 8/\log(1+4M)$. The optimization error is given by $r_T = f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^+_T)$, where $\boldsymbol{\theta}^*$ represents the global maximum point and $\boldsymbol{\theta}^+_T = \arg\max_{\boldsymbol{\theta}\in\mathcal{A}_T} f(\boldsymbol{\theta})$ denotes the approximation of the maximum point with the accumulated points $\mathcal{A}_T = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T\}$ from the previous T steps. Now, we have

$$r_T \leq \frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_t)) \tag{73}$$

$$\leq \frac{1}{T}\left(\sum_{t=1}^{T} 2\sqrt{\eta_t}\sigma_{t-1}(\boldsymbol{\theta}_t) + \sum_{t=1}^{T} 1/t^2\right) \tag{74}$$

$$\leq \frac{1}{T}\left(\sqrt{c_0 T\eta_T g_T} + \pi^2/6\right). \tag{75}$$

As $f(\boldsymbol{\theta})$ is considered as a sample drawn from a Gaussian process with $k_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, we can use the upper bound $\mathcal{O}(T^{\frac{p}{\nu+p}}\log^{\frac{\nu}{\nu+p}}(T))$ on the maximal g_T for $k_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in Ref. Vakili et al. (2021). By substituting η_T and $\mathcal{O}(T^{\frac{p}{\nu+p}}\log^{\frac{\nu}{\nu+p}}(T))$ into Eq. 75, the statement holds. □

Now we are ready to complete the proof of Theorem 4.5.

Proof of Theorem 4.5. We consider $\mathbb{V}[\partial_a f(\boldsymbol{\theta})]$ to be $1/\mathrm{poly}(n)$, as shown in Ref. Park & Killoran (2024). Additionally, we assume that the parameter dimension p is at most $\mathrm{poly}(n)$. To ensure consistency with the scenario under investigation and to guarantee the degree of discretization τ_t in Lemma C.2 of at least 1, we impose a constraint that the number of steps $T = \mathrm{poly}(n)$. Hence, it is reasonable to treat $\log\left(pT^2(\mathbb{V}_{\boldsymbol{\theta}}[\partial_a f(\boldsymbol{\theta})])^{1/2}\right)$ as a constant. Therefore, our task is to find the effective p that satisfies the condition $(p(\log(T)/T)^{\frac{\nu}{\nu+p}})^{1/2} \leq \epsilon$, where ϵ is a constant threshold and $T = \mathrm{poly}(n)$. Let

$$p \leq \frac{1}{2}\left(\epsilon^2 - \nu + \sqrt{(\epsilon^2-\nu)^2 + 4\nu\epsilon^2\left(1 + \log(T/\log T)\right)}\right), \tag{76}$$

then the above upper bound satisfies the inequality

$$p^2 - (\epsilon^2 - \nu)p - \nu\epsilon^2\left(1 + \log\left(T/\log T\right)\right) \leq 0. \tag{77}$$

Equivalently, the above inequality can be rewritten by

$$\log\left(\log T/T\right) \leq (1 + p/\nu)\left(1 - p/\epsilon^2\right). \tag{78}$$

Considering the relationship $\log x \geq 1 - 1/x$ holds for $x > 0$, then the above inequality implies

$$\log\left(\log T/T\right) \leq (1 + p/\nu)\log\left(\epsilon^2/p\right), \tag{79}$$

which directly leads to

$$\log T/T \leq \left(\epsilon^2/p\right)^{1+p/\nu}, \tag{80}$$

that is $(p(\log(T)/T)^{\frac{\nu}{\nu+p}})^{1/2} \leq \epsilon$. Finally, let $T = \mathrm{poly}(n^{1/\epsilon^2})$ and substitute it into Eq. 76. We obtain the effective parameter dimension p for the noise-free PS-QNN, which is $p \leq \tilde{\mathcal{O}}\left(\sqrt{\log n}\right)$. □

D Proof of Lemma 5.4

In this section, we provide the proof of Lemma 5.4 which is similar to the proof of Lemma 4.3.

Proof of Lemma 5.4. Given an n-qubit noisy PS-QNN objective function with q-strength local Pauli channels $\tilde{f}_q(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$, for any $j \in [2p]$, the partial derivatives $\partial_j \tilde{f}_q(\boldsymbol{\theta})$ exist and are continuous, as shown in Ref. Wang et al. (2021). Using a similar proof sketch as in Lemma B.2, we have

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D},\ \left|\tilde{f}_q(\boldsymbol{\theta}) - \tilde{f}_q(\boldsymbol{\theta}')\right| \leq \max_{j\in[2p]}\left(\sup_{\boldsymbol{\theta}\in\mathcal{D}}\left|\partial_j \tilde{f}_q(\boldsymbol{\theta})\right|\right)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1. \tag{81}$$

Considering the Maximum Cut problem on an unweighted d-regular graph with n vertices, we can rely on Corollary 2 in Ref. Wang et al. (2021) to obtain an upper bound on $\partial_j \tilde{f}_q(\boldsymbol{\theta})$ for any $j \in [2p]$. Then, the following relationship

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{D},\ \left|\tilde{f}_q(\boldsymbol{\theta}) - \tilde{f}_q(\boldsymbol{\theta}')\right| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 \tag{82}$$

holds, where the Lipschitz continuity factor is given by

$$L = \sqrt{\ln 2/2}d^2 n^{\frac{5}{2}}\|H_1^{\mathrm{MaxCut}}\|_\infty q^{((d_1+1)p+1)} \tag{83}$$

with the strength $q \in (0, 1)$ and d_1 representing the network depth of the implementation of the unitary corresponding to the problem-oriented Hamiltonian H_1^{MaxCut}. Since $\|H_1^{\mathrm{MaxCut}}\|_\infty = \mathcal{O}(nd/2)$, $q \in (0, 1)$ and $d_1 = \Omega(d)$, we have $L = \mathcal{O}(d^3 n^{7/2} q^{(d+1)p})$. Thus, the proof of Lemma 5.4 is concluded. □

E PROOF OF THEOREM 5.6

Theorem E.1 (Formal)**.** *Consider the Maximum Cut problem on an unweighted d-regular graph with n vertices, where d is a constant. Given a constant threshold ϵ, a failure probability $\delta \in (0,1)$ and a noisy PS-QNN objective function with q-strength local Pauli channels $\tilde{f}_q(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ induced by the network $\mathcal{U}_q(\boldsymbol{\theta})$ that satisfies Assumption 5.3, run BO for $T = \mathrm{poly}(n^{1/\epsilon^2})$ steps, where the scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as*

$$\eta_t = 2\log(\pi^2 t^2/(3\delta)) + 4p\log(4\pi p t^2 d^3 n^{7/2} q^{(d+1)p}). \tag{84}$$

Under the condition where the strength q spans $1/\mathrm{poly}(n)$ to $1/n^{1/\sqrt{\log n}}$, if the parameter dimension

$$p \leq \mathcal{O}\left(\log n / \log(1/q)\right), \tag{85}$$

then the optimization error $\tilde{r}_T$ satisfies $\tilde{r}_T \leq \epsilon$ with a success probability of at least $1-\delta$.

E.1 OUTLINE OF THE PROOF PROCEDURE

Our objective is to determine the effective parameter dimension p of the noisy PS-QNN $\mathcal{U}_q(\boldsymbol{\theta})$ such that the optimization error $\tilde{r}_T = \tilde{f}_q(\tilde{\boldsymbol{\theta}}^*) - \tilde{f}_q(\tilde{\boldsymbol{\theta}}_T^+)$ after $T = \mathrm{poly}(n)$ steps of executing BO can be upper bounded by a constant threshold ϵ. Here, $\tilde{\boldsymbol{\theta}}^*$ represents the global maximum point and $\tilde{\boldsymbol{\theta}}_T^+$ denotes the approximation of the maximum point in the previous T steps. We investigate this question through the perspective of the Bayesian approach, which considers the corresponding noisy PS-QNN objective function $\tilde{f}_q(\boldsymbol{\theta})$ as a sample drawn from a Gaussian process with the Matern covariance function $k_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$. We first establish that $\tilde{r}_T$ is upper bounded by $\frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_q(\tilde{\boldsymbol{\theta}}^*) - \tilde{f}_q(\tilde{\boldsymbol{\theta}}_t))$, where $\tilde{\boldsymbol{\theta}}_t$ represents the next point selected in each step t. It is evident that the condition $\frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_q(\tilde{\boldsymbol{\theta}}^*) - \tilde{f}_q(\tilde{\boldsymbol{\theta}}_t)) \leq \epsilon$ is sufficient to deduce the result $\tilde{r}_T \leq \epsilon$. Hence, by ensuring that the upper bound on $\frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_q(\tilde{\boldsymbol{\theta}}^*) - \tilde{f}_q(\tilde{\boldsymbol{\theta}}_t))$ is no greater than ϵ, we can determine the effective p that guarantees $\tilde{r}_T \leq \epsilon$. Subsequently, we utilize the continuity property of the noisy PS-QNN objective function $\tilde{f}_q(\boldsymbol{\theta})$ (Lemma 5.4) to establish an upper bound on $\frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_q(\tilde{\boldsymbol{\theta}}^*) - \tilde{f}_q(\tilde{\boldsymbol{\theta}}_t))$. The complete proof of Theorem 5.6 is similar to the proof of Theorem 4.5 and is supported by a series of lemmas analogous to Lemma C.2 to Lemma C.8. Instead of providing a detailed description of each lemma here, we will directly present lemma E.2 similar to Lemma C.8. Additionally, we will emphasize the impact of the difference in continuity property between the noise-free and noisy PS-QNN objective functions on the result.

E.2 PROOF DETAILS

In this section, we provide a comprehensive introduction to the Lemma E.2.

Lemma E.2. *Considering a Maximum Cut problem on an unweighted d-regular graph with n vertices, where d is a constant. Assuming that Assumption 5.3 holds, let $\tilde{f}_q(\boldsymbol{\theta}) : \mathcal{D} = [0, 2\pi]^{2p} \mapsto \mathbb{R}$ be the noisy PS-QNN objective function with q-strength local Pauli channels, where the strength $q \in (0,1)$. Given a failure probability $\delta \in (0,1)$, run BO with the Matern prior covariance function $k_{\mathrm{Matern}-\nu}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ (Eq. 23) for $T = \mathrm{poly}(n)$ steps, where a scaling parameter η_t for the acquisition function $\mathrm{UCB}_t(\boldsymbol{\theta})$ used in each step t is predefined as $\eta_t = 2\log(\pi^2 t^2/(3\delta)) + 4p\log(4\pi p t^2 d^3 n^{7/2} q^{(d+1)p})$. If the parameter dimension p is given by*

$$p \leq \mathcal{O}\left(\log n / \log(1/q)\right), \tag{86}$$

the optimization error $\tilde{r}_T$ satisfies

$$\tilde{r}_T \leq \mathcal{O}\left(\sqrt{p\log(pT^2 d^3 n^{7/2} q^{(d+1)p})(\log T/T)^{\frac{\nu}{\nu+p}}}\right) \tag{87}$$

with a success probability of at least $1-\delta$.

Proof. Using the continuity property of the noisy PS-QNN objective function $\tilde{f}_q(\boldsymbol{\theta})$ as stated in Lemma 5.4 and a series of lemmas similar to Lemma C.2 to Lemma C.8, we can easily obtain

the aforementioned result. It is essential to emphasize the constraint imposed on the parameter dimension p. To guarantee the degree of discretization τ_t of at least 1, as mentioned in Lemma C.2, we need to discuss the range of p that satisfies $pT^2d^3n^{7/2}q^{(d+1)p} \geq 1$. Since the number of steps $T = \mathrm{poly}(n)$ and p is at most $\mathrm{poly}(n)$, we can establish the inequality

$$n^{c_2} \leq pT^2n^{7/2} \leq n^{c_1}, \tag{88}$$

where c_1 and c_2 are two very close constants. Then, we have

$$\frac{p}{n^{c_1}d^3} \leq \frac{1}{T^2n^{7/2}d^3} \leq \frac{p}{n^{c_2}d^3}. \tag{89}$$

Since $q \in (0,1)$, the relationship

$$q^{\frac{p(d+1)}{n^{c_2}d^3}} \leq q^{\frac{d+1}{T^2n^{7/2}d^3}} \tag{90}$$

holds. As y^y is monotonically decreasing in the interval $(0, 1/e)$, we have

$$\left(\frac{1}{pT^2n^{7/2}d^3}\right)^{\frac{1}{pT^2n^{7/2}d^3}} \leq \left(\frac{1}{n^{c_1}d^3}\right)^{\frac{1}{n^{c_1}d^3}}. \tag{91}$$

Let

$$p \leq \frac{c_1 \log n + 3\log d}{(d+1)\log(1/q)n^{(c_1-c_2)}}, \tag{92}$$

then the above inequality implies

$$\left(\frac{1}{n^{c_1}d^3}\right)^{\frac{1}{n^{c_1}d^3}} \leq q^{\frac{p(d+1)}{n^{c_2}d^3}}, \tag{93}$$

which directly leads to

$$\left(\frac{1}{pT^2n^{7/2}d^3}\right)^{\frac{1}{pT^2n^{7/2}d^3}} \leq q^{\frac{d+1}{T^2n^{7/2}d^3}}, \tag{94}$$

that is $pT^2d^3n^{7/2}q^{(d+1)p} \geq 1$. Considering d as a constant, Eq. 92 implies $p \leq \mathcal{O}(\log n/\log(1/q))$. □

Proof of Theorem 5.6. Furthermore, when the strength $q \geq 1/\mathrm{poly}(n)$, it is reasonable to treat $\log(pT^2d^3n^{7/2}q^{(d+1)p})$ as a constant. Therefore, our objective is to determine the effective p that satisfies the condition $(p(\log(T)/T)^{\frac{\nu}{\nu+p}})^{1/2} \leq \epsilon$ with a constant threshold ϵ. The previous result shows that $p \leq \tilde{\mathcal{O}}(\sqrt{\log n})$ and $T = \mathrm{poly}(n^{1/\epsilon^2})$. Therefore, we have

$$p \leq \min\{\tilde{\mathcal{O}}(\sqrt{\log n}), \mathcal{O}(\log n/\log(1/q))\}. \tag{95}$$

Let $1/\mathrm{poly}(n) \leq q \leq 1/n^{1/\sqrt{\log n}}$, then this constraint implies

$$\log n/\log(1/q) \leq \sqrt{\log n}, \tag{96}$$

that is $p \leq \mathcal{O}(\log n/\log(1/q))$. Thus, the proof of Theorem 5.6 is concluded. □

F NUMERICAL EXPERIMENTS

We perform numerical experiments in three directions: (1) employing a more diverse set of graph structures, (2) comprehensively comparing BO and GD, and (3) numerically validating our theoretical results. To ensure conceptual clarity, we first define three key concepts for solving the Maximum Cut problems using PS-QNN: (1) Exact Solution of the Problem as the exact Maximum Cut value of a given graph, (2) Circuit-Achievable Value as the maximum objective function value attainable with the PS-QNN at specified depth, and (3) Algorithm-Optimized Value as the optimized objective function value obtained via BO or GD.

Table 2: Performance of BO on diverse graph structures.

Graph(Qubit=6)	1	2	3	4	5	6	7	8	9	10	Average
ExactSolution	5	5	7	8	6	6	7	8	6	8	6.6
AchievableValue	4.27	4.58	5.91	6.83	5.2	5.49	6.53	6.98	5.57	7.54	5.89
BO(Iteration=60)	4.19	4.41	5.69	6.62	5.14	5.35	6.2	6.84	5.25	7.12	5.68
Graph(Qubit=8)	1	2	3	4	5	6	7	8	9	10	Average
ExactSolution	7	9	9	11	12	13	13	11	11	16	11.2
AchievableValue	6.27	7.84	8.13	9.83	10.11	11.05	11.76	10.11	10.45	13.69	9.93
BO(Iteration=60)	5.98	7.61	7.47	9.38	9.66	10.65	11.3	9.75	10.14	12.37	9.43
Graph(Qubit=10)	1	2	3	4	5	6	7	8	9	10	Average
ExactSolution	16	14	14	12	14	18	18	16	22	24	16.8
AchievableValue	13.43	11.07	12.19	10.65	12.48	14.91	15.31	14.41	20.13	20.78	14.54
BO(Iteration=60)	12.5	10.8	11.65	10.37	11.85	14.66	14.42	13.78	19.16	19.89	13.91

Table 3: Performance comparison of BO and GD.

Graph(Qubit=10)	1	2	3	4	5	6	7	8	9	10	Average
ExactSolution	16	14	14	12	14	18	18	16	22	24	16.8
AchievableValue	13.43	11.07	12.19	10.65	12.48	14.91	15.31	14.41	20.13	20.78	14.54
BO(Iteration=30)	12.40	10.61	11.26	10.33	11.78	14.47	14.31	13.59	18.44	19.71	13.69
BO(Iteration=60)	12.50	10.80	11.65	10.37	11.85	14.66	14.42	13.78	19.16	19.89	13.91
BO(Iteration=90)	12.67	10.80	11.78	10.37	11.86	14.66	14.42	13.78	19.35	19.96	13.97
GD(Iteration=30)	12.35	9.50	10.84	9.71	11.16	12.01	13.27	11.94	18.43	18.70	12.79
GD(Iteration=60)	12.37	9.48	10.86	9.90	11.23	12.56	13.28	12.25	18.42	18.94	12.93
GD(Iteration=90)	12.55	9.47	10.87	9.92	11.39	12.59	13.28	12.26	18.47	18.97	12.98

F.1 Performance of BO on diverse graph structures

We investigate random graphs with 6, 8, and 10 vertices (10 graphs per size) and construct the Maximum Cut objective function using a 2-layer PS-QNN. For each graph, we run BO with 10 random initializations and 60 iterations per run. The results, summarized in Table 2, demonstrate that BO performs robustly, achieving average accuracies-defined as the ratio of the mean Algorithm-Optimized Value to the mean Exact Solution-of 86.06%, 84.20%, and 82.80% for graphs with 6, 8, and 10 vertices, respectively.

F.2 Performance Comparison of BO and GD

We comprehensively compare the performance of BO and GD by evaluating the Algorithm-Optimized Value and the number of steps to convergence. This comparison uses 10 randomly generated 10-vertex graphs, with the Maximum Cut objective function constructed for each graph using a 2-layer PS-QNN. To ensure a rigorous comparison, BO and GD are executed with 10 random initializations and tested for 30, 60, and 90 iterations. The results are summarized in Table 3.

Table 4: Numerical validation of theoretical results.

Graph(Qubit=10)	1	2	3	4	5	6	7	8	9	10	Average
ExactSolution	16	14	14	12	14	18	18	16	22	24	16.8
AchievableValue(depth=1)	12.46	9.73	11.13	9.91	11.73	13.54	14.02	13.45	18.90	20.21	13.51
Iteration(ϵ=0.9)	2	3	4	6	9	3	6	9	13	17	7.2
Iteration(ϵ=0.8)	3	4	4	7	9	4	6	9	19	18	8.3
Iteration(ϵ=0.7)	3	4	6	7	9	5	6	11	21	22	9.4
AchievableValue(depth=2)	13.43	11.07	12.19	10.65	12.48	14.91	15.31	14.41	20.13	20.78	14.54
Iteration(ϵ=2.0)	15	7	13	1	5	11	13	10	26	9	11
Iteration(ϵ=1.5)	19	12	17	8	14	19	18	20	36	18	18.1
Iteration(ϵ=1.0)	36	16	28	15	18	25	31	25	59	43	29.6

F.3 Numerical Validation of Theoretical Results

Recognizing that error mitigation techniques can effectively address quantum circuit noise, we focus our analysis on the noiseless scenario. Our experiments use 10 randomly generated 10-vertex graphs, with the Maximum Cut objective function implemented via 1-layer and 2-layer PS-QNNs. For the 1-layer PS-QNN, we analyze the relationship between the optimization error ϵ-defined as the difference between Circuit-Achievable Value and Algorithm-Optimized Value-and average iteration counts T at error levels of 0.7, 0.8, 0.9. Similarly, for the 2-layer PS-QNN, we examine this relationship at error levels of 1, 1.5, 2. In both cases, we observe $\log T \propto 1/\epsilon^2$. The detailed results are summarized in Table 4.

G The use of large language models(LLMs)

During the preparation of this work, we use LLMs to assist in language polishing and editing the initial draft. This tool is used solely to improve grammatical fluency and sentence structure.