

Optimal algorithmic complexity of inference in quantum kernel methods [1]

Elies Gil-Fuster,^{1,2} Seongwook Shin,^{1,3} Sofiene Jerbi,^{1,4} Jens Eisert,^{1,2,4} and Maximilian J. Kramer¹

¹*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany*

²*Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*

³*Department of Quantum Information Science and Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, 05006, Seoul, Republic of Korea*

⁴*Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany*

Quantum kernel methods [2] are among the top candidates for quantum advantage in *quantum machine learning* (QML), as showcased in the milestone work of Ref. [3]. Typical QML tasks involve a data domain $\mathcal{X}$ and a label co-domain $\mathcal{Y}$, and the goal is to obtain a function $f: \mathcal{X} \rightarrow \mathcal{Y}$ which performs well under pre-specified metrics. The mainstream approach is based on *embedding quantum kernels* [4–6], where a kernel function k is specified via an n -qubit *quantum feature map* $x \mapsto \rho(x)$, for $x, x' \in \mathcal{X}$, as $k(x, x') = \text{Tr}\{\rho(x)\rho(x')\}$. A crucial step in any kernel-based pipeline is related to *training*, i.e., finding suitable parameters $\alpha \in \mathbb{R}^N$ from a training set of size N . After training, one must evaluate the resulting hypothesis on new data – the *inference* step – which amounts to computing $f_\alpha(x) = \sum_{i=1}^N \alpha_i k(x, x_i)$.

Commonly used techniques [7] estimate each of the N terms $\alpha_i k(x, x_i)$ independently via sampling-based amplitude estimation, resulting in a runtime of $\mathcal{O}(N\|\alpha\|_2^2/\epsilon^2)$, with $\|\alpha\|_2^2 = \sum_{i=1}^N \alpha_i^2$. We refer to this approach as *list-and-sum*. In this submission, we propose two independent axes of improvement: In the way how individual kernel values are estimated (*via sampling* versus *via quantum amplitude estimation*), and how the sum over training points is approximated (*list-and-sum* versus *all-at-once*). We sketch this framework in Fig. 1.

Results: We show that in the early fault-tolerant regime [8], the optimal combination – encoding the full sum as a single observable and applying quantum amplitude estimation – achieves a query complexity of $\mathcal{O}(\|\alpha\|_1/\epsilon)$, where $\|\alpha\|_1 = \sum_{i=1}^N |\alpha_i|$. We complement this with a matching lower bound of $\Omega(\|\alpha\|_1/\epsilon)$, establishing optimality up to logarithmic factors. Beyond query complexity, our results provide a practical guide to quantum kernel inference: given hardware capabilities, we provide a meta-algorithm and meta-theorem which identify the appropriate strategy and its runtime, as well as their corresponding gate complexity. We establish that the strategy yielding the lowest total gate count is achieved *list-and-sum* with adaptive budget exploiting quantum amplitude estimation.

Main ideas: We work in the oracle query complexity model mainly to be able to establish a meaningful comparison between shot-based algorithms and those based on quantum amplitude estimation. We introduce a reasonable access model to several quantum oracles for the main purpose of deriving worst-case statements: we characterize the complexity of inference in quantum kernel methods without exploiting any structure. The intuitive idea for the two improvements can be drawn directly from the definition of the problem: $f_\alpha(x)$ is a *weighted* sum over several quantities that correspond to the *expectation value of a Bernoulli-distributed random variables*. On the one hand, being careful with norms and signs, mathematically one can always re-write a weighted sum as an expectation value. Our technical contribution is to provide an efficient algorithm to measure the *all-at-once* observable associated to this re-writing. On the other hand, the machinery of quantum amplitude estimation was natively designed precisely to estimate expectation values of Bernoulli-distributed random variables (encoded in quantum states). Our technical contribution is to generically reduce the problem of inference to that of estimating a single amplitude, and provide efficient algorithms to instantiate the reduction. One expects a generic quadratic improvement from quantum amplitude estimation, we show that we obtain a second improvement (from $N\|\alpha\|_2^2$ to $\|\alpha\|_1$) on top.

Potential impact: Our results should be deemed relevant already from a purely computer-science perspective, inference is a necessary step in any QML pipeline using kernel methods, and its complexity had not been characterized before in our practically-minded access model. By providing a matching lower bound, we completely settle this question within our access model. More broadly, our study hints at a potential historical difference between classical and quantum ML. Deep neural networks have been state-of-the-art in classical ML for longer than a decade, and classical kernel methods are often deemed impractical due to their runtime scaling in the number of training data N . Roughly speaking, training a neural network runs in time $\sim \mathcal{O}(N)$, and inference for neural networks runs in time $\sim \mathcal{O}(1)$; training a usual kernel method in turn runs in time $\sim \mathcal{O}(N^2)$, and inference in kernel methods runs in time $\sim \mathcal{O}(N)$. The relevant difference within this submission is the gap between $\mathcal{O}(1)$ (neural networks) and $\mathcal{O}(N)$ (kernel methods). By showing that the complexity of inference in quantum kernel methods does not depend explicitly on N , we show that quantum kernel methods could remain practical as the number of training data increases.

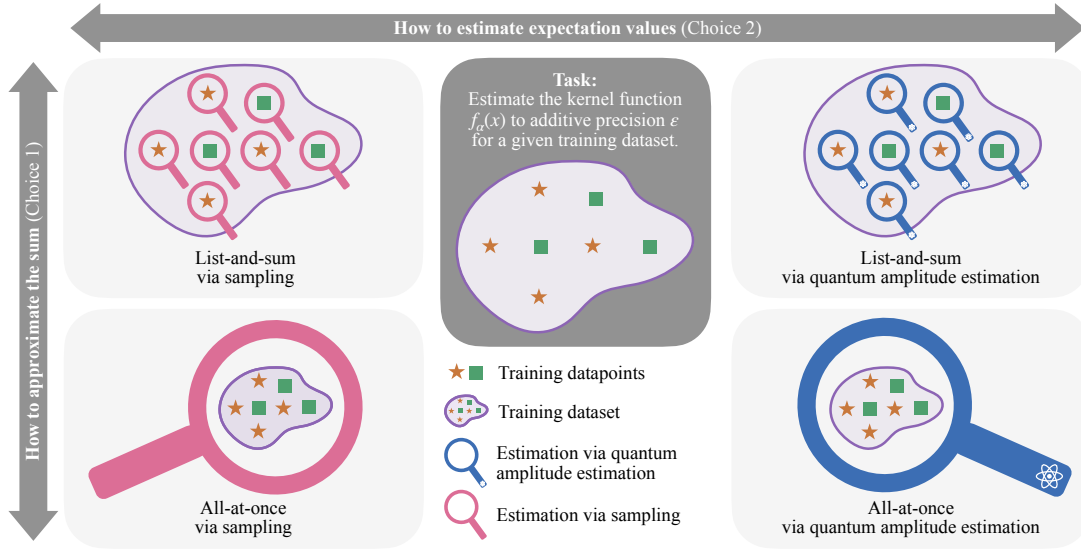


FIG. 1. Sketch of algorithmic improvements.

I. Inference in Quantum Kernel Methods. Suppose a trained quantum kernel model $f_\alpha(x) = \sum_{i=1}^N \alpha_i k(x, x_i)$, with trained parameters $\alpha \in \mathbb{R}^N$, given a training set $S = \{(x_i, y_i)\}_{i=1}^N$ of size N . Here, the kernel function is defined from a parametrized unitary $x \mapsto U(x)$ as $k(x, x') = |\langle 0|U^\dagger(x)U(x')|0\rangle|^2$. The goal is to provide a quantum algorithm which, for any new input x , evaluates the kernel function $f_\alpha(x)$ to additive precision ϵ with probability at least $2/3$.

We separate the coefficient vector into three parts: (1) its norm $\|\alpha\|_1$, (2) a probability distribution $p = (p_i)_{i=1}^N$ with $p_i = |\alpha_i|/\|\alpha\|_1$, and (3) a vector of signs $s = (s_i)_{i=1}^N$, with $s_i = \text{sign}(\alpha_i)$. With this, we can write $f_\alpha(x) = \|\alpha\|_1 \sum_{i=1}^N p_i s_i k(x, x_i) = \|\alpha\|_1 \mathbb{E}_{i \sim p}[s_i k(x, x_i)]$. In parallel, accounting for the positive and negative terms in the sum, we also introduce the intermediate variables $f_\pm \in [0, 1]$ as $f_\pm := \sum_{i=1}^N \delta_{s_i, \pm 1} p_i k(x, x_i)$. We can re-write $f_\alpha(x)$ in terms of these normalized quantities as $f_\alpha(x) = \|\alpha\|_1 (f_+ - f_-)$.

II. Amplitude Estimation. Let $V \in \mathbb{C}^{d \times d}$ be a unitary such that $V|0 \dots 0\rangle|0\rangle = a|\psi_{\text{good}}\rangle|0\rangle + \sqrt{1-a^2}|\psi_{\text{bad}}\rangle|1\rangle$, where $|\psi_{\text{good}}\rangle$ and $|\psi_{\text{bad}}\rangle$ are normalized quantum state vectors, and $a \in [0, 1]$. Given access to V and $V^\dagger$, the task is to output an estimate $\hat{a}$ such that $a - \epsilon < \hat{a} < a + \epsilon$ with probability at least $2/3$.

Two approaches are relevant: *sampling*, which estimates a from repeated measurements of $V|0\rangle$ with sampling complexity $\mathcal{O}(\epsilon^{-2})$, and *quantum amplitude estimation* (QAE), which uses coherent applications of V and $V^\dagger$ to achieve $\mathcal{O}(\epsilon^{-1})$ [9, 10]. The latter scaling is optimal [11] and requires inverse access [12, 13].

III. Access Model. We consider the following quantum oracles and specify the associated gate costs per query. (1) *Single-input oracle* $U(x)$: given any input x and an n -qubit register, we can implement $U(x)$, $U^\dagger(x)$, and controlled versions thereof, with an associate cost of G -many gates. (2) *Coefficient quantum-state oracle* $W(\alpha)$: Given a $(\lceil \log_2(N) \rceil + 1)$ -qubit register, prepare a quantum state encoding the coefficients as amplitudes $|\alpha\rangle = \sum_{i=1}^N \sqrt{p_i}|i\rangle \otimes |s_i\rangle$, and s_i encoding the sign of the i^{th} coefficient as an eigenstate of the Z operator as $Z|s_i\rangle = s_i|s_i\rangle$. The associated cost per query in terms of gates is N .

We note that the gate cost per query varies across algorithms, yet in a way that is mostly independent of ϵ and $\|\alpha\|_1$, and hence it does not affect the asymptotic comparisons in runtime and query complexity we perform in this work. The access model we propose allows us to perform a formal analysis of query complexity, and it also accurately reflects realistic implementations. The oracles we introduce can be directly instantiated on hardware, and they do not require any special computational power, nor do they hide exponential overheads. Still, we assume *black-box* access to $U(x)$ and $W(\alpha)$ to remove the possibility of optimized implementations which would exploit the structure of the circuits involved. This is a conscious design choice, as it is our goal to focus on the fundamental complexity of the computational task, which in turn provides worst-case guarantees applicable to any specific quantum feature map. As is common in query complexity studies, nothing prevents more-efficient procedures from existing under the assumption that $U(x)$ is of a particular known form.

III. All at once observable. Instead of estimating each kernel term individually, we reinterpret the weighted sum as a single expectation value by embedding the coefficients into a quantum state. This allows us to evaluate the entire inference expression in one coherent procedure, rather than aggregating many separate estimates. We start by introducing another quantum oracle that can be constructed from those in our access model.

Training-set oracle $O^\dagger(S)$. Given two registers, an n -qubit “data” register and a $\lceil \log_2(N) \rceil$ -qubit “index” register, which indexes each element $i \in \{1, \dots, N\}$, we call $O^\dagger(S) := \sum_{i=1}^N U^\dagger(x_i)_{\text{data}} \otimes |i\rangle\langle i|_{\text{idx}}$ the *training-set oracle*. We note that this construction does not require *quantum random-access memory* (QRAM). The $\mathcal{O}(N)$ gate cost per query is inherent to any procedure – classical or quantum – that must touch all N training points.

Lemma 1 (Query complexity training-set-oracle). *The training-set oracle can be realized with N controlled queries to the single-input oracle $U(x_i)$. The associated gate cost is $\mathcal{O}(N(G + \log N))$, and the procedure uses one extra qubit.*

Proof. The associated gate cost is, for each $i \in \{1, \dots, N\}$: $\mathcal{O}(\log N)$ gates to perform a multi-controlled bitflip on an extra flag register conditioned on the index being i , $\mathcal{O}(G)$ to call the controlled version of the G gates in $U(x_i)$, and $\mathcal{O}(\log N)$ to uncompute the flag bitflip, for a total of $\mathcal{O}(G + \log N)$. Since these must be repeated N times, the total gate cost of the training-set oracle is $\mathcal{O}(N(G + \log N))$. $\square$

Theorem 2 (All-at-once observable). *There exists a bounded Hermitian observable $\mathcal{M}(\alpha, S)$, with $\|\mathcal{M}(\alpha, S)\|_\infty \leq \|\alpha\|_1$, and a quantum state vector $|\phi(x)\rangle$, such that $\langle \phi(x) | \mathcal{M}(\alpha, S) | \phi(x) \rangle = f_\alpha(x)$. Moreover, this expectation value can be implemented using $n + \lceil \log_2 N \rceil + 2$ qubits and $\mathcal{O}(N(G + \log N))$ gates.*

Proof sketch. Set $\mathcal{M}(\alpha, S) = \sum_{i=1}^N \alpha_i |\phi(x_i)\rangle\langle \phi(x_i)|$ and verify directly. The resource estimate comes from Lemma 1. $\square$

IV. Resource estimates. The two algorithmic choices (*sampling* vs *quantum amplitude estimation*, and *list-and-sum* versus *all-at-once*) can be combined freely, giving rise to a landscape of algorithms. For each combination, we prove the resulting runtime complexity; the results are collected in terms of query complexity in Theorem 3 and in terms of gates in Theorem 4.

Theorem 3 (Query complexity of inference in quantum kernel methods). *There exist algorithms for estimating $f_\alpha(x)$ to additive precision ε with constant success probability. The algorithmic runtime depends on the algorithmic choices as follows:*

Strategy	Sampling	Quantum amplitude estimation
1(a): list-and-sum, fixed	$\mathcal{O}\left(\frac{N\ \alpha\ _2^2}{\varepsilon^2}\right)^\ddagger$	$\mathcal{O}\left(\frac{N\ \alpha\ _2}{\varepsilon}\right)$
1(b): list-and-sum, adaptive	$\mathcal{O}\left(\frac{\ \alpha\ _1^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\ \alpha\ _{2/3}}{\varepsilon}\right)^\diamond$
1(c): all-at-once	$\mathcal{O}\left(\frac{\ \alpha\ _1^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\ \alpha\ _1}{\varepsilon}\right)^\star$

We highlight $\ddagger$ as the naive approach (1a, sampling), $\star$ as query-optimal (1c, quantum amplitude estimation), and $\diamond$ as gate-optimal (1b, quantum amplitude estimation; see Theorem 4).

In terms of runtime complexity, Theorem 3 dictates that the best algorithmic choice depends on the available hardware. On near-term devices without coherent multi-query access, using an adaptive budget for *list-and-sum* already yields the same runtime scaling as measuring the more-costly *all-at-once* observable. Conversely, the *all-at-once* approach is more beneficial in the early-fault-tolerant regime, where the runtime scaling is directly manifested by the 1-norm of the vector of coefficients, versus the $2/3$ -norm achieved by *list-and-sum* with an adaptive budget. Explicitly, within any case for Choice 1, quantum amplitude estimation yields a quadratic improvement over sampling. We expand beyond query-complexity analysis in Theorem 4, where we show the asymptotic scaling of the number of gates associated to each choice.

Theorem 4 (Number of gates for inference in quantum kernel methods). *There exist algorithms for estimating $f_\alpha(x)$ to additive precision ε with constant success probability. The number of gates depends on the algorithmic choices as follows:*

Strategy	Sampling	Quantum amplitude estimation
1(a): list-and-sum, fixed	$\mathcal{O}\left(\frac{GN\ \alpha\ _2^2}{\varepsilon^2}\right)^\ddagger$	$\mathcal{O}\left(\frac{GN\ \alpha\ _2}{\varepsilon}\right)$
1(b): list-and-sum, adaptive	$\mathcal{O}\left(\frac{G\ \alpha\ _1^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{G\ \alpha\ _{2/3}}{\varepsilon}\right)^\diamond$
1(c): all-at-once	$\tilde{\mathcal{O}}\left(\frac{NG\ \alpha\ _1^2}{\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{(NG+n)\ \alpha\ _1}{\varepsilon}\right)^\star$

We highlight $\ddagger$ as the naive approach (1a, sampling), $\star$ as query-optimal (1c, quantum amplitude estimation; see Theorem 3), and $\diamond$ as gate-optimal (1b, quantum amplitude estimation).

Practical guidelines. For practitioners, our results suggest the following guidelines for choosing an inference strategy:

- Use *all-at-once* combined with quantum amplitude estimation when coherent multi-query access is available, as this achieves query-optimal scaling.
- Use *list-and-sum* with *adaptive budget* combined with quantum amplitude estimation when gate cost dominates, as this minimizes the total number of gates.
- Sampling-based approaches are primarily competitive in near-term regimes without coherent access, but are asymptotically suboptimal in precision.

-
- [1] E. Gil-Fuster, S. Shin, S. Jerbi, J. Eisert, and M. J. Kramer, [arXiv: 2604.15214](#) (2026).
 - [2] J. Tanner, C.-F. Kam, and J. Wang, arXiv preprint arXiv:2604.07896 (2026), [arXiv:2604.07896](#).
 - [3] Y. Liu, S. Arunachalam, and K. Temme, *Nature Phys.* **17**, 1013 (2021).
 - [4] M. Schuld, arXiv preprint arXiv:2101.11020 (2021), [arXiv:2101.11020](#).
 - [5] T. Hubregtsen, D. Wierichs, E. Gil-Fuster, P.-J. H. S. Derks, P. K. Faehrmann, and J. J. Meyer, *Phys. Rev. A* **106**, 04243 (2022).
 - [6] E. Gil-Fuster, J. Eisert, and V. Dunjko, *Machine Learning: Science and Technology* **5**, 025003 (2024).
 - [7] V. Havlíček *et al.*, *Nature* **567**, 209 (2019).
 - [8] J. Eisert and J. Preskill, arXiv preprint arXiv:2510.19928 (2025), [arXiv:2510.19928](#).
 - [9] G. Brassard, P. Høyer, M. Mosca, and A. Tapp, *Quantum Computation and Information*, 53–74 (2002).
 - [10] S. Aaronson and P. Rall, Quantum approximate counting, simplified, in *Symposium on Simplicity in Algorithms* (Society for Industrial and Applied Mathematics, 2020) p. 24–32.
 - [11] A. Nayak and F. Wu, in *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99 (Association for Computing Machinery, New York, NY, USA, 1999) p. 384–393.
 - [12] E. Tang and J. Wright, arXiv preprint arXiv:2507.23787 (2025), [arXiv:2507.23787](#).
 - [13] E. Tang, J. Wright, and M. Zhandry, arXiv preprint arXiv:2510.07622 (2025), [arXiv:2510.07622](#).