

Quantum Maximum Likelihood Prediction via Hilbert Space Embeddings

Sreejith Sreekumar and Nir Weinberger

L2S, CentraleSupélec, CNRS, University of Paris-Saclay and Technion, Israel

Full version of paper available at [arXiv:2602.18364](https://arxiv.org/abs/2602.18364) [1]

Introduction: Prediction is a primitive task with numerous applications in machine learning, statistics, information theory, physics, finance, and engineering. This problem has been studied from multiple perspectives, e.g., machine learning [2] and information theory [3]. In the classical theory, the performance of the predictor degrades with the dimension of the parameter vector used to determine the predictor. Specifically, under the common log loss (cross-entropy), the expected regret in predicting that a sample from a probability distribution P follows a probability distribution Q is given by the Kullback–Leibler (KL) divergence, $D_{\text{KL}}(P\|Q)$. In the minimax framework, the regret scale linearly with this dimension of the model class of Q (e.g., [4–7], [8, Sec. 13.4]). However, in modern applications, the parameter dimensionality could be large, e.g., (i) the number of parameters in the architecture of a deep neural network, or any of the various refinements of their effective dimension (e.g. [9]), or (ii) the cardinality of the vocabulary of token sequences in large language models (LLMs). Despite this seemingly large effective dimension, modern prediction algorithms work remarkably well in practice, suggesting that training and prediction must rely on some inductive bias, whether explicit or implicit.

Here, we assume that the inductive bias is captured by embedding the letters x of the alphabet $\mathcal{X}$ (discrete or a subset of $\mathbb{R}^m$) into vectors in a Hilbert space $\mathbb{H}$ of dimension d . Let $\varphi : \mathcal{X} \rightarrow \mathbb{H}$ be a given embedding such that it is Borel-measurable and $\|\varphi(x)\|_{\mathbb{H}} = 1$ for all $x \in \mathcal{X}$. For a probability measure P with density p with respect to a dominating positive σ -finite measure μ (say counting measure in the discrete case and Lebesgue measure in the continuous case), set $\rho_p = \int_{\mathcal{X}} p(x) |\varphi(x)\rangle \langle \varphi(x)| d\mu(x)$. The *covariance embedding* [10] is the linear map ϕ induced by $P \mapsto \rho_p$, which maps a probability distribution to a covariance operator, or to a density operator in quantum parlance [11, 12]. Given a set Σ of density operators and samples $x^n \in \mathcal{X}^n$ generated according to an unknown distribution $P^{\otimes n}$ with empirical distribution $\hat{P}_n$, the *quantum maximum likelihood predictor* (QMLP) is defined as any minimizer of

$$\inf_{\sigma \in \Sigma} D(\hat{\rho}_n(x^n) \|\sigma), \quad \text{with} \quad \hat{\rho}_n(x^n) = \phi(\hat{P}_n), \quad (1)$$

whenever the infimum is attained, where $D(\rho \|\sigma) := \text{Tr}[\rho(\log \rho - \log \sigma)]$ is the quantum relative entropy between quantum states ρ and σ . A classical predictor over $\mathcal{X}$ can be obtained by processing the QMLP to an output probability distribution (in the case of a quantum system, this involves applying a measurement, i.e., a positive operator-valued measure (POVM), according to the Born rule). This induces the bias that closeness in quantum relative entropy correlates with predictive accuracy on the output alphabet.

Intuitively, since the embedding is to a finite-dimensional Hilbert space of dimension much smaller than the vocabulary, the curse of large vocabulary is alleviated, and the discrete geometry between words in the vocabulary (Hamming distance or “zero-one” similarity) is replaced by the inner product in the embedded space. A prominent representative of this idea are LLMs, in the sense that during inference they induce a probability distribution over the next token (symbol, word, sentence, etc.) based on previous ones. Such a viewpoint provides a geometry in which token similarity is quantified by inner products and hence provides a natural setting where optimization algorithms such as stochastic gradient descent can be performed efficiently and analyzed.

Main Contributions: As our main contribution, we establish performance guarantees for QMLP summarized as follows. We obtain non-asymptotic upper bounds (Theorem 1) quantifying the expected rate of convergence of QMLP (and MLP) to the target state both in trace norm and quantum relative entropy. We also show the corresponding concentration inequalities. These bounds depend on the dimension d of the Hilbert space, the sample size n , and the minimal positive eigenvalue of the embedded quantum state corresponding to the true data distribution. For instance, our bounds show that the rate of convergence in trace norm is $\tilde{O}(d/\sqrt{n})$ (up to an additive approximation slack factor) in general and

$\tilde{O}(d^3/n)$ when the class of quantum models is sufficiently expressive, where $\tilde{O}(\cdot)$ hides logarithmic factors and dependence on minimal eigenvalue. Under additional conditions on the expressivity of quantum models measured in terms of quantum relative entropy, the corresponding rates of convergence under quantum relative entropy are $\tilde{O}(d/\sqrt{n})$ and $\tilde{O}(d^2/n)$, respectively. We also derive upper bounds on the covering number of quantum states under relative entropy.

Background and Related Work: Embedding probability distributions via mean embedding into Hilbert space has been explored in a variety of other contexts [13–18]. The covariance embedding [10] of probability distributions into density operators has also found several applications in terms of defining new divergence measures [19], proposing kernel-based likelihood test [20] and unsupervised learning [21]. The covariance embedding of probability distributions and the associated QMLP also captures at a high-level the problem of prediction in quantum LLMs [22], where the expressivity of quantum architecture used for prediction is reflected in the class of models over which QMLP is optimized. Another related topic corresponds to quantum generative modeling and simulation of probability distributions, in which the hope is to use quantum models to efficiently simulate probability distributions which are difficult to model classically. Two popular frameworks used in this regard pertains to quantum circuit Born machines [23, 24] and quantum Boltzmann machines [25–27], which uses quantum circuits and thermal (or Gibbs) states, respectively, to achieve the modeling objective.

Key Technical Contributions: For technical reasons, we will consider a slightly perturbed version of $\hat{\rho}_n(x^n)$ obtained using the maximally mixed state π_d ($:= I/d$ where I denotes the identity operator) in $\mathbb{H}$ such that the perturbation vanishes asymptotically with n . Accordingly, define

$$\rho_n(x^n) := \left(1 - \frac{1}{n}\right) \hat{\rho}_n(x^n) + \frac{1}{n} \pi_d, \quad (2)$$

and set

$$\hat{\sigma}_n^*(x^n) := \left(1 - \frac{1}{n}\right) \sigma_n^*(x^n) + \frac{1}{n} \pi_d \quad \text{with} \quad \sigma_n^*(x^n) := \arg \min_{\sigma \in \Sigma} D(\rho_n(x^n) \parallel \sigma). \quad (3)$$

The next theorem establishes consistency and convergence properties of QMLP.

Theorem 1 Let $\rho_p > 0$ and $\rho_n(x^n) > 0$ be density operators on $\mathbb{H}$ as defined above. Suppose Σ is a compact set with $\text{supp}(\Sigma) = \mathbb{H}$ such that $\sigma_p^* := \arg \min_{\sigma \in \Sigma} D(\rho_p \parallel \sigma)$ satisfies $D(\rho_p \parallel \sigma_p^*) \leq \epsilon$. Set

$$b_n := \log \|\sigma_p^{*-1}\| \vee \log(dn) \vee T(\rho_p, \sigma_p^*) \quad \text{and} \quad \bar{b}_n := \log(dn) \vee \log \|\rho_p^{-1}\|. \quad (4)$$

Then, with $\sigma_n^*(x^n)$ and $\hat{\sigma}_n^*(x^n)$ as defined in (3), the following hold for $X^n \sim P^{\otimes n}$:

(i) **Performance guarantees under (squared) trace norm:**

$$\mathbb{E}[\|\sigma_n^*(X^n) - \rho_p\|_1^2] \leq \begin{cases} 16d(b_n + 4)(\sqrt{2 \log(2d)} + \sqrt{\pi/2})n^{-\frac{1}{2}} \\ \quad + 16b_n n^{-1} + 64n^{-2} + 28\epsilon, & \text{if } \epsilon > 0, \\ (32d \|\rho_p^{-1}\| + 8d^2) \left(\frac{8}{n^2} + \frac{28d}{n}\right), & \text{if } \epsilon = 0. \end{cases} \quad (5)$$

Moreover, for all $t \geq 0$, we have

$$\mathbb{P}\left(\|\sigma_n^*(X^n) - \rho_p\|_1^2 \geq \frac{16b_n}{n} + \frac{64}{n^2} + 28\epsilon + 8d(b_n + 4)t\right) \leq 4de^{-\left(\frac{nt^2}{4} \wedge \frac{3nt}{4}\right)}, \quad \text{if } \epsilon > 0, \quad (6a)$$

$$\mathbb{P}\left(\|\sigma_n^*(X^n) - \rho_p\|_1^2 \geq (32d \|\rho_p^{-1}\| + 8d^2) \left(\frac{8}{n^2} + 2t\right)\right) \leq 2de^{-\left(\frac{nt}{4} \wedge \frac{3n\sqrt{t}}{4}\right)}, \quad \text{if } \epsilon = 0. \quad (6b)$$

(ii) **Performance guarantees under quantum relative entropy:** If Σ is such that

$$\mathbb{E} [D(\rho_n(X^n) \|\sigma_n^*(X^n))] \wedge \mathbb{E} [D(\rho_n(X^n) \|\hat{\sigma}_n^*(X^n))] \leq \epsilon, \quad (7)$$

then

$$\mathbb{E} [D(\rho_p \|\hat{\sigma}_n^*(X^n))] \leq \begin{cases} (2\bar{b}_n + \log d)n^{-1} + d\bar{b}_n(\sqrt{2\log(2d)} + \sqrt{\pi/2})n^{-\frac{1}{2}} + \epsilon, & \text{if } \epsilon > 0, \\ \frac{28d^2\|\rho_p^{-1}\| + 2\log d}{n} & \text{if } \epsilon = 0. \end{cases} \quad (8)$$

If Σ also satisfies

$$D(\rho_n(X^n) \|\sigma_n^*(X^n)) \wedge D(\rho_n(X^n) \|\hat{\sigma}_n^*(X^n)) \leq \epsilon, \quad \text{almost surely,} \quad (9)$$

then for all $t \geq 0$,

$$\mathbb{P} \left(D(\rho_p \|\hat{\sigma}_n^*(X^n)) \geq \frac{2\bar{b}_n}{n} + \frac{\log d}{n} + \epsilon + d\bar{b}_n t \right) \leq 2de^{-\left(\frac{nt^2}{4} \wedge \frac{3nt}{4}\right)}, \quad \text{if } \epsilon > 0, \quad (10a)$$

$$\mathbb{P} \left(D(\rho_p \|\hat{\sigma}_n^*(X^n)) \geq 2dt \|\rho_p^{-1}\| + \frac{2\log d}{n} \right) \leq 2de^{-\left(\frac{nt}{4} \wedge \frac{3n\sqrt{t}}{4}\right)}, \quad \text{if } \epsilon = 0. \quad (10b)$$

The proof of Theorem 1 relies on utilizing the variational form of quantum relative entropy [28, 29] to control $\|\sigma_n^*(X^n) - \rho_p\|_1$ and $D(\rho_p \|\hat{\sigma}_n^*(X^n))$, and then applying matrix Bernstein or Hoeffding bounds. Observe that (5) implies a parametric rate of convergence with respect to n for the QMLP to the target distribution (up to the approximation factor ϵ) when $\epsilon > 0$ and $O_d(1/n)$ rate of convergence when $\epsilon = 0$. Hence, there is no curse of dimensionality in the convergence rate. Also, note that (8) and (10), combined with Pinsker's inequality, yield the analogues of (5) and (6), respectively, with $\|\sigma_n^*(X^n) - \rho_p\|_1$ replaced by $\|\hat{\sigma}_n^*(X^n) - \rho_p\|_1$. Further, by data processing inequality for trace norm and quantum relative entropy, the claims in Theorem 1 also hold for the output probability distributions obtained after processing the QMLP and the embedded target distribution ρ_p .

To obtain a set Σ satisfying the conditions in Theorem 1, it suffices to take any ϵ -covering of the set $\mathcal{S}$ of all density operators in quantum relative entropy, i.e., a set $\mathcal{A}$ of quantum states such that for any $\rho \in \mathcal{S}$, there exists a $\sigma \in \mathcal{A}$ with $D(\rho \|\sigma) \leq \epsilon$. Let $N(\epsilon, \mathcal{S}, D(\cdot \|\cdot))$ denote the minimum cardinality of such a set $\mathcal{A}$, which is referred to as the covering number of $\mathcal{S}$ in quantum relative entropy. The next proposition provides an upper bound on this quantity.

Proposition 1 There exists $0 < c < 68$ such that

$$N(\epsilon, \mathcal{S}, D(\cdot \|\cdot)) \leq (cd\epsilon^{-1})^{2d^2+d-1}, \quad \forall \quad 0 < \epsilon \leq 1. \quad (11)$$

Future Outlook: Our performance guarantees for QMLP indicate that a good embedding should aim for minimal embedded Hilbert space dimension and maximal smallest eigenvalue for the embedded density operator. While this provides some insights into the choice of an efficient embedding, often it is desirable for embeddings to have additional properties such as injectivity or approximate relative distance preserving property within specific sub-classes of probability distributions. Hence, it is of interest to characterize embeddings and distribution classes for which such properties hold. Another aspect pertains to relaxing the assumption of independent and identically distributed samples to incorporate Markov or other kinds of dependence. Yet another avenue to explore pertains to reverse contraction coefficients for specific embeddings which quantifies how much the quantum relative entropy in the embedded domain is preserved compared to KL divergence in the distribution space (see the recent study by [30]). Finally, development of efficient algorithms for solving the QMLP optimization problem along with associated theoretical guarantees is of considerable interest; see [31] for classical results and [32] for recent developments in the framework of quantum Boltzmann machines.

References

- [1] S. Sreekumar and N. Weinberger, “Quantum maximum likelihood prediction via Hilbert space embeddings,” 2026. [Online]. Available: <https://arxiv.org/abs/2602.18364>
- [2] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.
- [3] N. Merhav and M. Feder, “Universal prediction,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.
- [4] R. Krichevsky and V. Trofimov, “The performance of universal encoding,” *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.
- [5] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1983.
- [6] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals of Statistics*, pp. 1564–1599, 1999.
- [7] D. Haussler and M. Opper, “Mutual information, metric entropy and cumulative relative entropy risk,” *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.
- [8] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.
- [9] N. Golowich, A. Rakhlin, and O. Shamir, “Size-independent sample complexity of neural networks,” *Information and Inference: A Journal of the IMA*, vol. 9, no. 2, pp. 473–504, 2020.
- [10] F. Bach, “Information theory with kernel methods,” *IEEE Transactions on Information Theory*, vol. 69, no. 2, pp. 752–775, 2023.
- [11] M. M. Wilde, *Quantum Information Theory*, 2nd ed. Cambridge University Press, 2017.
- [12] M. Hayashi, *Quantum Information Theory: Mathematical Foundation*. Springer Berlin, Heidelberg, 2016.
- [13] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA, USA: Kluwer Academic Publishers, 2004.
- [14] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, “Kernel mean embedding of distributions: A review and beyond,” *Foundations and Trends in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.
- [15] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, “A kernel method for the two-sample-problem,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, ser. NIPS’06. Cambridge, MA, USA: MIT Press, 2006, p. 513–520.
- [16] K. Fukumizu, F. R. Bach, and M. I. Jordan, “Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces,” *Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.
- [17] A. Smola, A. Gretton, L. Song, and B. Schölkopf, “A Hilbert space embedding for distributions,” in *Algorithmic Learning Theory*, M. Hutter, R. A. Servedio, and E. Takimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 13–31.
- [18] L. Song, A. Gretton, and C. Guestrin, “Nonparametric tree graphical models,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterton, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 765–772.

- [19] J. K. Hoyos-Osorio and L. G. Sanchez-Giraldo, “The representation Jensen-Shannon divergence,” 2024.
- [20] L. V. Santoro and V. M. Panaretos, “Likelihood ratio tests by kernel Gaussian embedding,” arXiv preprint, Aug 2025.
- [21] O. Kachaiev and S. Recanatesi, “Learning to embed distributions via maximum kernel entropy,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 44 710–44 734, 2024.
- [22] I. Basile and F. Tamburini, “Towards quantum language models,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1840–1849.
- [23] J.-G. Liu and L. Wang, “Differentiable learning of quantum circuit born machines,” *Phys. Rev. A*, vol. 98, p. 062324, Dec 2018.
- [24] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, “A generative modeling approach for benchmarking and training shallow quantum circuits,” *npj Quantum Information*, vol. 5, no. 1, p. 45, 2019.
- [25] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, “Quantum-assisted learning of hardware-embedded probabilistic graphical models,” *Phys. Rev. X*, vol. 7, p. 041052, Nov 2017.
- [26] M. Kieferová and N. Wiebe, “Tomography and generative training with quantum Boltzmann machines,” *Phys. Rev. A*, vol. 96, p. 062327, Dec 2017.
- [27] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, “Quantum Boltzmann machine,” *Phys. Rev. X*, vol. 8, p. 021050, May 2018.
- [28] D. Petz, “A variational expression for the relative entropy,” *Communications in Mathematical Physics*, vol. 114, no. 2, pp. 345–349, 1988.
- [29] M. Berta, O. Fawzi, and M. Tomamichel, “On variational expressions for quantum relative entropies,” *Letters in Mathematical Physics*, vol. 107, no. 12, pp. 2239–2265, Dec. 2015.
- [30] P. Belzig, L. Gao, G. Smith, and P. Wu, “Reverse-type data processing inequality,” *Communications in Mathematical Physics*, vol. 406, no. 12, p. 295, 2025.
- [31] I. Csiszár and P. C. Shields, “Information theory and statistics: A tutorial,” *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.
- [32] M. M. Wilde, “Fundamentals of quantum Boltzmann machine learning with visible and hidden units,” *arXiv: 2512.19819*, 2025.