

Sampling from Thermal Quantum States via Pauli Propagation

Michele Minervini,^{1,2} Josh Goldman,¹ Mark M. Wilde,² Zoë Holmes,^{1,3} and Manuel S. Rudolph^{1,3}

¹*Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

²*School of Electrical and Computer Engineering,
Cornell University, Ithaca, New York 14850, USA*

³*Centre for Quantum Science and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Pauli propagation is a powerful framework for simulating quantum systems by tracking the evolution of Pauli operators, yet its scope has so far been limited to estimating expectation values. We extend it to sampling from quantum states, introducing an algorithm that draws bitstrings directly from truncated Pauli expansions. While the very truncation that makes Pauli propagation efficient renders the represented operator non-positive-semidefinite – so that its Born “probabilities” can be negative and standard sampling is ill-defined – we resolve this with a locally normalized conditional sampling procedure that builds a valid distribution on the fly, bypassing the intractable global renormalization over all 2^n bitstrings. We prove that the total variation distance to the true Born distribution scales linearly with the truncation error, and that the cost per sample is linear in the number of retained Pauli terms. We deploy the sampler on thermal states prepared by imaginary-time Pauli propagation – a natural setting, since the evolution starts from the identity, the sparsest Pauli operator. Because quantum Boltzmann machines (QBMs) are parameterized thermal states, this combination further enables their training for generative learning tasks; numerically, we demonstrate both accurate sampling from truncated thermal states and successful training of these models. More broadly, our method offers a new pathway for simulating, sampling from, and training models of complex quantum systems on classical hardware.

Introduction — Simulating the measurement of a quantum state ρ in the computational basis is a central task in quantum information: the goal is to produce bitstrings $x = x_1 \cdots x_n \in \{0, 1\}^n$ distributed according to the Born rule, $p(x) = \text{Tr}[\Pi_x \rho]$ with $\Pi_x = |x\rangle\langle x|$. This task also underlies many proposed routes to quantum advantage, notably in quantum generative modeling, which rest precisely on the hardness of producing such samples classically [1, 2]. Essentially all classical sampling algorithms perform this task by drawing the bits sequentially, a strategy known as *ancestral sampling* [3, 4]: the j -th bit is drawn from the conditional distribution

$$p(x_j | x_{<j}) = \frac{p(x_{\leq j})}{p(x_{<j})}, \quad p(x_{<j}) = \text{Tr}[\Pi_{x_{<j}} \rho], \quad (1)$$

where $x_{<j} := x_1 \cdots x_{j-1}$ is the already-sampled prefix, $x_{\leq j}$ is that prefix extended by the candidate bit x_j , and $\Pi_{x_{<j}} = |x_{<j}\rangle\langle x_{<j}| \otimes I$ projects onto outcomes consistent with the prefix. Sampling is thereby assembled from marginal-probability evaluations [5–8], so the practical question is which classical representations of the state ρ allow these marginals to be evaluated efficiently and reliably. Beyond restricted circuit families [9–11], the leading classical representations used for sampling have been tensor networks [12, 13], which struggle for highly entangled two-dimensional systems [14], and neural quantum states [15, 16], which sample natively but lack error guarantees [17].

Pauli propagation is a well-established alternative for classical simulation [18–20]: it tracks the coefficients of an operator expanded in the Pauli basis. While originally proposed for simulating in the Heisenberg picture [18, 19], recently it has been shown that high-temperature thermal states can be simulated via Pauli propagation in the Schrödinger picture [21, 22][23]. In this case, one ends up with a Pauli sum representation of the state of the form $\rho = 2^{-n} \sum_P c_P P$. This representation is kept efficient by *truncating* the expansion during the propagation – discarding strings with small coefficients or large Pauli weight (the number of qubits on which a string acts nontrivially) – so that only a retained set $\mathcal{S}$ of $M := |\mathcal{S}|$ terms is kept, $\tilde{\rho} = 2^{-n} \sum_{P \in \mathcal{S}} c_P P$.

Sampling from truncated Pauli sums — **So far Pauli propagation’s scope has been confined to estimating expectation values; here we establish it as a sampler.** The obstacle is that the truncated operator $\tilde{\rho}$ is generally no longer a valid quantum state: it fails to be positive semidefinite, so the quantities entering the chain rule (1) become *quasi-probabilities* $q(x) := \text{Tr}[\Pi_x \tilde{\rho}]$, which can be negative and can also exceed unity. Ratios of quasi-marginals then no longer define valid conditional distributions, and ancestral sampling breaks down. This loss of a valid probability model under an otherwise accurate approximation is a sign problem, akin to those in quantum Monte Carlo [24] and negative quasi-probability representations [25]; the present submission overcomes this obstruction.

The most natural fix is to take absolute values and renormalize once, globally: define the induced distribution $\tilde{p}(x) := |q(x)|/\mathcal{N}$ with $\mathcal{N} = \sum_y |q(y)|$. This candidate is statistically sound: we prove that $\text{TV}(p, \tilde{p}) \leq \|\rho - \tilde{\rho}\|_1$, where TV denotes the total variation distance (TVD). Yet it is computationally intractable. The absolute value breaks the linearity by which a marginal is the trace of a reduced operator: a marginal of $\tilde{p}$ can only be obtained by explicitly summing $|q|$ over all the completions of the prefix, and the normalization $\mathcal{N}$ itself is a sum over all 2^n

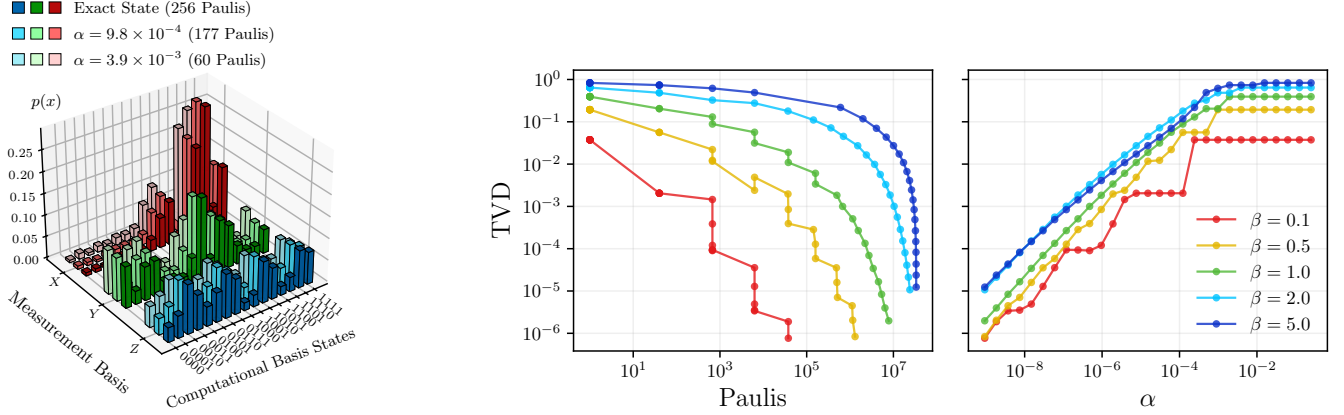


FIG. 1. Sampling from truncated thermal states. Left: distributions sampled from a 4-qubit, $\beta = 10$ thermal state of a random two-body Hamiltonian, in the Z, Y, and X bases: exact state (256 Pauli terms) versus truncated operators with 177 and 60 terms, obtained by discarding all coefficients below the thresholds $\alpha = 2^{-10}$ and 2^{-8} , respectively. Right: TVD between sampled and exact thermal distributions for a 14-qubit Heisenberg chain, versus retained Pauli strings and truncation threshold α , across inverse temperatures β .

bitstrings. The sequential structure of Eq. (1), however, suggests where to intervene: ancestral sampling explores a binary tree whose depth- j nodes are the prefixes $x_{<j}$, and instead of normalizing once at the root, as the global law does, we normalize at every node. For each step j we define the locally normalized conditional probability

$$\hat{p}(x_j | x_{<j}) := \frac{|q(x_{<j}, x_j)|}{|q(x_{<j}, 0)| + |q(x_{<j}, 1)|}, \quad \hat{p}(x) := \prod_{j=1}^n \hat{p}(x_j | x_{<j}), \quad (2)$$

where the two terms in the denominator are the quasi-marginals of the two candidate values, 0 and 1, of the next bit, each computed as $q(x_{<j}) = \text{Tr}[\Pi_{x_{<j}} \tilde{\rho}]$ directly from the retained coefficients $\{c_P\}_{P \in \mathcal{S}}$. Each step requires two quasi-marginal evaluations instead of exponentially many, the sign problem is eliminated locally, and if $\tilde{\rho} \succeq 0$, $\hat{p}$ reduces exactly to the Born distribution.

The cost per sample is set by the quasi-marginal evaluations. Only retained strings acting as I or Z on every measured qubit contribute to $q(x_{<j})$: X and Y have zero diagonal, so any string acting off-diagonally on a measured qubit traces to zero. Scanning at most M retained strings at each of the n steps gives a worst-case cost $O(nM)$ per sample; for generic expansions, whose strings carry roughly random Pauli letters, a string survives the measurement of a new qubit only if its letter there is I or Z – which happens with probability about one half – so the number of contributing strings roughly halves at every step and the expected total cost drops to $O(M)$; thus sampling is as efficient as the propagation that produced the expansion. Moreover, the same representation also serves as an exact likelihood evaluator: the model probability $\hat{p}(x)$ of any given bitstring can be computed *pointwise exactly* in $O(nM)$ time by multiplying the n conditionals in Eq. (2) – a combination rare among generative samplers, which we exploit below for likelihood-based training.

Error guarantees — The accuracy of the sampler is inherited from the quality of the truncation, quantified by the trace-norm error $\varepsilon := \|\tilde{\rho} - \rho\|_1$.

Theorem (TVD bound for locally normalized sampling). *Let ρ be an n -qubit state with Born distribution p , let $\tilde{\rho}$ be its truncated approximation with truncation error $\varepsilon = \|\tilde{\rho} - \rho\|_1$, and let $\hat{p}$ be the sampled distribution defined in Eq. (2). Then $\text{TV}(p, \hat{p}) \leq 2\varepsilon$.*

We further isolate the role of the non-physicality of the truncated operator: writing $\{\lambda_i\}_i$ for the eigenvalues of $\tilde{\rho}$, the spectral negativity $N(\tilde{\rho}) := \sum_{\lambda_i < 0} |\lambda_i|$ measures the weight of the negative eigenvalues created by the truncation, and we prove that, whenever the smallest prefix marginal $s_{\min}$ encountered during sampling exceeds $4N(\tilde{\rho})$, $\text{TV}(p, \hat{p}) \leq \varepsilon + 2nN(\tilde{\rho})/(s_{\min} - 4N(\tilde{\rho}))$: the first term is the intrinsic approximation error, the second quantifies how negativity destabilizes the n local normalizations. Finally, the coefficient bound $\|\rho - \tilde{\rho}\|_1 \leq (\sum_{P \notin \mathcal{S}} c_P^2)^{1/2}$ connects these guarantees to the truncation rule used in practice: the sampling error is controlled by the squared weight of the *discarded* coefficients, so discarding the smallest coefficients is precisely the choice that minimizes the error.

Sampling thermal states prepared by imaginary-time propagation — Since the cost per sample is linear in M , the method calls for states whose Pauli representation is naturally sparse. Thermal states $\rho_\beta = e^{-\beta H} / \text{Tr}[e^{-\beta H}]$ are a direct such case: writing them as the two-sided imaginary-time evolution (ITE) of the identity, $\rho_\beta \propto e^{-\beta H/2} I e^{-\beta H/2}$, they are reached starting from a single Pauli string, with the operator growing as β increases – so the approach scales

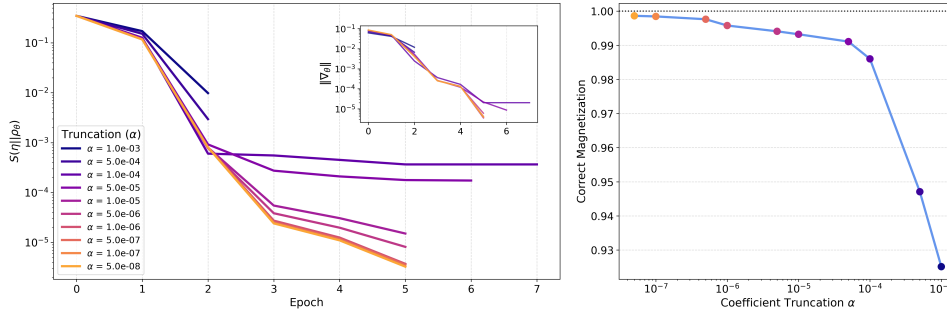


FIG. 2. End-to-end QBM training: a 20-qubit QBM learns a transverse-field Ising thermal target ($J = 1$, $h = 1.2$, $\beta = 0.5$). Left: relative-entropy loss versus epoch for thresholds $\alpha \in [5 \times 10^{-8}, 10^{-3}]$; inset: gradient norm. Right: accuracy of the learned transverse magnetization versus α .

best in the high-temperature regime, while small systems remain accessible at any temperature. For $H = \sum_k h_k G_k$ with Pauli strings G_k , Trotterization [26, 27] splits $e^{-\beta H/2}$ into r steps of the elementary factors $e^{-\theta G_k}$, $\theta = h_k \beta / (2r)$, each updating a Pauli string P by the two-sided rule

$$e^{-\theta G} P e^{-\theta G} = \begin{cases} P & \text{if } P \text{ and } G \text{ anticommute,} \\ \cosh(2\theta) P - \sinh(2\theta) PG & \text{if } P \text{ and } G \text{ commute.} \end{cases} \quad (3)$$

Figure 1 demonstrates the combined preparation-and-sampling procedure: for a 4-qubit low-temperature ($\beta = 10$) thermal state of a random two-body Hamiltonian, compressing the exact 256-term expansion down to 177 and even 60 terms yields non-physical operators whose sampled distributions still capture the physical features of the exact state, in the computational as well as the rotated X and Y measurement bases (left panel). At larger scale the convergence is systematic: for a 14-qubit Heisenberg chain, the TVD to the exact thermal distribution decreases monotonically with the number of retained strings and with the truncation threshold, across temperatures (right panel).

Training QBMs — A direct application of preparing and sampling thermal states in this way is generative modeling with quantum Boltzmann machines (QBMs) [28, 29]. The model distribution is the computational-basis diagonal $p_\theta(x) = \langle x | \rho_\theta | x \rangle$ of a parameterized Gibbs state $\rho_\theta \propto e^{-\beta H_\theta}$, with $H_\theta = \sum_k \theta_k H_k$. When H_θ is diagonal, the model reduces to the classical Boltzmann machine [30, 31]; non-commuting terms generate coherences, and hence correlations, inaccessible to the classical model with the same parameter budget [28]. The training loss depends on the target: for a quantum target η , one minimizes the quantum relative entropy $S(\eta || \rho_\theta)$, whose gradient $\partial_{\theta_k} S = \text{Tr}[H_k \eta] - \text{Tr}[H_k \rho_\theta]$ [29] requires expectation values on the current Gibbs model [32]; for classical data, one instead minimizes the Kullback–Leibler divergence between the empirical data distribution and p_θ . Our framework supplies all the required primitives from one truncated Pauli representation: expectation values (read off the coefficients), exact pointwise likelihoods $\hat{p}_\theta(x)$ (entering the classical loss and its gradients), and samples (generation).

Figure 2 shows end-to-end training of a 20-qubit QBM on a transverse-field Ising thermal target. The relative-entropy loss converges smoothly for all ten truncation thresholds (left panel), and the trained model’s transverse magnetization tracks the exact value, degrading from $\sim 99.9\%$ accuracy at the smallest α to $\sim 93\%$ at the largest (right panel). Furthermore, Figure 3 shows images generated by a (heavily truncated) Pauli propagation Boltzmann machine with 30 qubits. It was trained with the Kullback–Leibler divergence, which we can exactly evaluate between the quantum-inspired model and the data set. Preliminary results show that the learned models capture the dominant structure of the data and generate faithful low-resolution digits at scales beyond typical QBM studies. Scaling to many more qubits while keeping images sharp remains challenging, leaving room for algorithmic innovation and co-development with quantum hardware.

Outlook — We have introduced a sampling algorithm for truncated Pauli expansions (a locally normalized chain-rule construction valid for non-positive operators) with cost linear in the retained terms and TVD guarantees linear in the truncation error. This can be viewed as a generative-modeling engine with exact likelihoods, certified sampling error, and a classical baseline for quantum-advantage claims. Although the cost of the method is linear in the number of terms, this number will typically grow polynomially with system size, potentially with a large polynomial prefactor. By contrast, sampling from the quantum device is comparatively inexpensive and essentially size-independent. Thus, while our algorithm is efficient in the usual polynomial-time sense, one should still expect polynomial speedups from direct quantum sampling.

Here we have illustrated the method for thermal states. However, it applies to any approximation of a state with a sparse Pauli representation, including operators that are not strictly quantum states (only a non-zero trace is required), and, beyond the Pauli basis, to any quasi-probability representation that develops negativity [25]. Thus we believe this method has much broader applicability than shown here and we leave exploring these for future work.

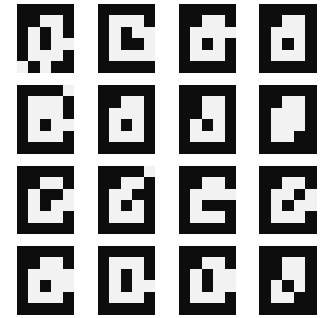


FIG. 3. Eight down-sampled MNIST images and eight generated images by a 6×5 Pauli propagation Boltzmann machine. Take a guess.

-
- [1] M. J. Bremner, R. Jozsa, and D. J. Shepherd, Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **467**, 459–472 (2010).
 - [2] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, *Nature* **574**, 505–510 (2019).
 - [3] S. Bravyi, D. Gosset, and Y. Liu, How to simulate quantum measurement without computing marginals, *Physical Review Letters* **128** (2022).
 - [4] A. J. Ferris and G. Vidal, Perfect sampling with unitary tensor networks, *Physical Review B* **85**, 10.1103/physrevb.85.165146 (2012).
 - [5] H. Pashayan, S. D. Bartlett, and D. Gross, From estimation of quantum probabilities to simulation of quantum circuits, *Quantum* **4**, 223 (2020).
 - [6] M. Van Den Nest, Classical simulation of quantum computation, the Gottesman–Knill theorem, and slightly beyond, *Quantum Information and Computation* **10**, 258–271 (2010).
 - [7] R. Jozsa, Classical simulation and complexity of quantum computations, in *Computer Science – Theory and Applications* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010) pp. 252–258.
 - [8] E. Bernstein and U. Vazirani, Quantum complexity theory, in *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, STOC ’93 (Association for Computing Machinery, New York, NY, USA, 1993) p. 11–20.
 - [9] D. Gottesman, The Heisenberg representation of quantum computers, *arXiv preprint quant-ph/9807006* (1998).
 - [10] L. G. Valiant, Quantum circuits that can be simulated classically in polynomial time, *SIAM Journal on Computing* **31**, 1229 (2002).
 - [11] B. M. Terhal and D. P. DiVincenzo, Classical simulation of noninteracting-fermion quantum circuits, *Physical Review A* **65** (2002).
 - [12] G. Vidal, Efficient simulation of one-dimensional quantum many-body systems, *Physical Review Letters* **93** (2004).
 - [13] M. S. Rudolph and J. Tindall, *Simulating and sampling from quantum circuits with 2d tensor networks* (2025), *arXiv:2507.11424* [quant-ph].
 - [14] F. Pan, K. Chen, and P. Zhang, Solving the sampling problem of the Sycamore quantum circuits, *Physical Review Letters* **129** (2022).
 - [15] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602–606 (2017).
 - [16] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, *Physical Review Letters* **124** (2020).
 - [17] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, *Nature Communications* **9** (2018).
 - [18] P. Rall, D. Liang, J. Cook, and W. Kretschmer, Simulation of qubit quantum circuits via Pauli propagation, *Physical Review A* **99** (2019).
 - [19] M. S. Rudolph, T. Jones, Y. Teng, A. Angrisani, and Z. Holmes, *Pauli propagation: A computational framework for simulating quantum systems* (2025), *arXiv:2505.21606* [quant-ph].
 - [20] T. Begušić, J. Gray, and G. K.-L. Chan, Fast and converged classical simulations of evidence for the utility of quantum computing before fault tolerance, *Science Advances* **10** (2024).
 - [21] M. S. Rudolph, A. Angrisani, A. Wright, I. Sanderski, R. Puig, and Z. Holmes, *Thermal state simulation with pauli and majorana propagation* (2026), *arXiv:2602.04878* [quant-ph].
 - [22] R. Gómez-Lurbe and A. Pérez, *Pauli Propagation for Imaginary Time Evolution* (2026), *arXiv:2601.14400* [quant-ph].
 - [23] Our submission is strongly aligned with the submission entitled ‘Thermal State Simulation with Pauli and Majorana Propagation’[21]. We think that, together, they represent a notable contribution to QML, and the program committee might consider merging them.
 - [24] M. Troyer and U.-J. Wiese, Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations, *Physical Review Letters* **94** (2005).
 - [25] V. Veitch, C. Ferrie, D. Gross, and J. Emerson, Negative quasi-probability as a resource for quantum computation, *New Journal of Physics* **14**, 113011 (2012).
 - [26] H. F. Trotter, On the product of semi-groups of operators, *Proceedings of the American Mathematical Society* **10**, 545 (1959).
 - [27] M. Suzuki, Generalized Trotter’s formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems, *Communications in Mathematical Physics* **51**, 183 (1976).
 - [28] H. J. Kappen, Learning quantum models from quantum or classical data, *Journal of Physics A: Mathematical and Theoretical* **53**, 214001 (2020).
 - [29] M. Kieferova and N. Wiebe, Tomography and generative data modeling via quantum Boltzmann training, *Physical Review A* **96**, 062327 (2017), *arXiv:1612.05204* [quant-ph].
 - [30] G. E. Hinton and T. J. Sejnowski, Optimal perceptual inference, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 1983) pp. 448–453.
 - [31] R. Salakhutdinov and G. E. Hinton, Deep Boltzmann machines, in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)* (PMLR, 2009) pp. 448–455.
 - [32] L. Coopmans and M. Benedetti, On the sample complexity of quantum boltzmann machine learning, *Communications Physics* **7** (2024).