15

Gibbs sampling

The authors are grateful to Rolando Somma for reviewing this chapter.

Rough overview (in words)

Gibbs sampling is the task of preparing a quantum state in thermal equilibrium. This task is interesting in its own right as a means of testing the thermodynamic properties of quantum systems in a controlled way, but it is also a subroutine that is surprisingly useful within other quantum algorithms. Formally, given a Hamiltonian and a temperature, the task is to prepare the *Gibbs state* (also known as the *thermal state*) of that Hamiltonian at the associated temperature, or equivalently, to sample eigenstates of the Hamiltonian with probability proportional to their Boltzmann weights (motivating the name Gibbs *sampling*).

Physically, Gibbs sampling is routinely achieved in experiments via cooling as a manifestation of open-system thermodynamics, although theoretical understanding of such processes has been largely heuristic. Computationally, quantum Gibbs sampling is the quantum analog of the same classical task in the computational basis, often achieved by Markov chain Monte Carlo (MCMC) methods. As a representative example, the Metropolis–Hastings algorithm [512] performs local accept-reject steps to construct a Markov chain whose stationary state is the classical Gibbs distribution; the Gibbs distribution can be efficiently sampled if the Markov chain mixes rapidly. Nowadays, Monte Carlo methods have already outgrown their original intent and found ubiquitous applications in optimization and machine learning due to their simplicity and versatility. It is natural to wonder if the same features will be present for quantum Gibbs sampling.

The most direct quantum algorithms for Gibbs sampling (for noncommuting Hamiltonians) suffer from an explicit cost exponential in the size of the system. Another approach is to quantize classical Monte Carlo algorithms [984], but this approach has faced serious technical challenges rooted in quantum mechanics: the energy-time uncertainty principle (for imposing the Boltzmann weights) and the no-cloning theorem (for "rejecting" a quantum state). Recently, a new wave [256, 939, 856, 260, 259] of proposals revisits the issue from the angle of open-system thermodynamics and gives nature-inspired algorithms for Gibbs sampling. These more directly resemble the dynamical process of thermalization and have the potential to achieve better runtimes for specific systems where thermalization is expected to be fast.

Rough overview (in math)

Given a Hamiltonian $H = \sum_i E_i |\psi_i\rangle\langle\psi_i|$ over n qubits, a desired inverse temperature β, and an error parameter ϵ, the Gibbs sampling task is to prepare an n-qubit quantum state ρ such that

$$\|\rho - \sigma_\beta\|_{\mathrm{tr}} \leq \epsilon\,,$$

where

$$\sigma_\beta := \frac{\mathrm{e}^{-\beta H}}{\mathcal{Z}} \propto \sum_i \mathrm{e}^{-\beta E_i} |\psi_i\rangle\langle\psi_i| \quad \text{and} \quad \mathcal{Z} := \mathrm{tr}[\mathrm{e}^{-\beta H}].$$

The quantity $\mathcal{Z}$ is known as the partition function. The above uses the convenient error metric given by the trace norm $\|\cdot\|_{\mathrm{tr}}$, which controls the error for arbitrary bounded (possibly nonlocal) observables. In some applications, it could be sufficient to give a state ρ that approximates all *local* observables up to high precision, even if the global distance between ρ and σ_β is large. Note that σ_β corresponds to an ensemble of eigenstates of H, where an eigenstate with energy E_i occurs with probability proportional to the Boltzmann weight $\mathrm{e}^{-\beta E_i}$ (β has units of inverse energy so that βE_i is dimensionless).

To solve this problem, the quantum algorithm requires access to H, for example, through a block-encoding of H. Block-encodings can often be efficiently constructed, for instance, when H is a sparse matrix or when H is given as a sum of poly(n) local interaction terms. Henceforth, assume that H is offset such that it is guaranteed to be a non-negative operator (no negative energies).

An early approach [838] for Gibbs sampling relied on quantum phase estimation (QPE) and amplitude amplification. In particular, one starts with a $2n$-qubit maximally entangled state (for which the reduced density matrix on the first n qubits is the maximally mixed state) and applies QPE to the first n qubits, reading an estimate of the energy into an ancilla register. Under the simplification that QPE has perfect resolution, one now has the state

$$\frac{1}{\sqrt{2^n}} \sum_i |\psi_i\rangle|\phi_i\rangle|E_i\rangle\,,$$

where $|\psi_i\rangle$ is the i-th eigenstate of H, E_i is the associated energy, and the states $|\phi_i\rangle$ form an arbitrary (unimportant) orthonormal basis. Next, one coherently rotates an ancilla qubit to put the correct Boltzmann weight into the amplitude:

$$\frac{1}{\sqrt{2^n}} \sum_i |\psi_i\rangle|\phi_i\rangle|E_i\rangle\Big(\mathrm{e}^{-\beta E_i/2}|0\rangle + \sqrt{1-\mathrm{e}^{-\beta E_i}}|1\rangle\Big)\,.$$

Note that the probability of measuring the final qubit in $|0\rangle$ is precisely $\mathcal{Z}/2^n$. Rather than measure and postselect, one now performs amplitude amplification on the ancilla being $|0\rangle$ to produce

$$\frac{1}{\sqrt{\mathcal{Z}}} \sum_i \mathrm{e}^{-\beta E_i/2}|\psi_i\rangle|\phi_i\rangle|E_i\rangle$$

up to small error, which is a purification of the Gibbs state $\sigma_\beta = \mathcal{Z}^{-1}\sum_i \mathrm{e}^{-\beta E_i}|\psi_i\rangle\langle\psi_i|$. While QPE does not exactly produce the operation described above, a more complete analysis in [838, 272] shows the idea still works. This approach is akin to classical rejection sampling (see also [822]), where a state is chosen at random and accepted with probability $\mathrm{e}^{-\beta E_i}$, such that repeating until acceptance yields a sample from the correct distribution. Due to amplitude amplification, the quantum algorithm enjoys a quadratic speedup.

More advanced methods that have exponentially better ϵ dependence have since been developed. Reference [289] used a linear combination of unitaries approach to perform the imaginary-time evolution operator $\mathrm{e}^{-\beta H}$, again followed by amplitude amplification. Technically, that work assumed access to an operator similar to $\sqrt{H}$, but this requirement was removed in Gibbs samplers appearing in [48, 45], which employ a method for implementing smooth Hamiltonian functions. Alternatively, one can use the quantum singular value transformation along with a polynomial approximation to the function $\mathrm{e}^{-\beta(1-x)/2}$ on the interval $x \in [-1, 1]$ [431, Section 5.3] and combine this with (fixed-point) amplitude amplification [1067].

Another family of quantum algorithms is closer in spirit to classical Monte Carlo methods. They quantize the Metropolis–Hastings algorithm (quantum Metropolis sampling [984]) or simulate the dynamics arising from a system-bath interaction [256, 939, 856, 260, 259]. These algorithms make fundamental usage of QPE (or the quantum operator Fourier transform [260, 259]) for probing the energy, but most importantly (and most nontrivially), they construct a detailed-balance "quantum Markov chain" via either discretely or continuously "rejecting" the quantum state. Intuitively, the algorithm implements the following rule:

> "Apply a jump. If the energy increases, reject with some probability."

Care must be taken to perform the rejection step coherently and to handle the fact that the energies cannot be learned to infinite precision. Abstractly, Monte Carlo–style quantum algorithms emulate a continuous-time quantum Markov chain (i.e., a Lindbladian $\mathcal{L}$)[1] that converges to the Gibbs state after evolution time t_{mix} (often called the *mixing time*)

$$\mathcal{L}[\sigma_\beta] \approx 0 \quad \text{and} \quad \|e^{\mathcal{L}t_{\text{mix}}}[\rho_0] - \sigma_\beta\|_{\text{tr}} \leq \epsilon$$

for some initial state ρ_0. Like the classical Metropolis–Hastings algorithm, for some systems, t_{mix} can be exponentially large (or worse) in the system size n, while for other systems, it can be much smaller (polynomial or logarithmic). It is a generally difficult problem to determine t_{mix}. As a consequence, classical Monte Carlo algorithms rarely have convergence guarantees in practice. Rather, they are employed in conjunction with a variety of heuristic convergence tests.

Note that such a process can be further quantized to gain quadratic speedup [1048, 260].

Dominant resource cost (gates/qubits)

Assuming one has access to a block-encoding of the Hamiltonian H, that is, a unitary whose upper left block is the operator H/α, where α is a normalization constant at least as large as the spectral norm of H, one can accomplish the Gibbs sampling task using [48, Lemma 44] (see also [45, Corollary 16])

$$\alpha\beta\sqrt{\frac{2^n}{\mathcal{Z}}} \cdot \text{poly}(\log(1/\epsilon), n) \tag{15.1}$$

calls to the block-encoding and a similar number of other gates. Note that since we have assumed H is non-negative, we have $\mathcal{Z} \leq 2^n$. In the case that H is d-sparse, we can take $\alpha = d\|H\|_{\max}$, where $\|H\|_{\max}$ is the maximum absolute value of any entry of the matrix H. In the case one has access to $\sqrt{H}$, the β dependence can be reduced from β to $\sqrt{\beta}$ [289]. This complexity statement might be regarded as a *quadratic speedup* compared to the classical method of rejection sampling, which requires $2^n/\mathcal{Z}$ samples on average; however, note that this classical method only directly applies to diagonal (classical) Hamiltonians H.

[1] Most constructions are continuous-time quantum Markov chains as they are mathematically easier to work with than discrete-time quantum channels [984].

Otherwise, a classical approach may need to resort to exact diagonalization of H, which has $\mathcal{O}(2^n)$ space complexity and even worse time complexity.

Monte Carlo–style quantum Gibbs sampling algorithms have complexity determined by

$$(\text{mixing time}) \cdot (\text{cost per unit evolution time } \mathrm{e}^{\mathcal{L}}). \tag{15.2}$$

The mixing time is expected to vary significantly for different systems of interest (based on classical Monte Carlo intuition), but for systems appearing in nature, one may be optimistic based on the observed fast thermalization of physically relevant observables. In fact, in many cases, rapidly thermalizing metastable states will have physically relevant characteristics; in analogy with classical ferromagnetic systems below the critical temperature. The cost per unit evolution time is dominated by the QPE subroutine, which then scales with a certain energy resolution. Typically, we may work in a convention where H is given by poly(n) Hamiltonian terms each of strength $\mathcal{O}(1)$, enabling a block-encoding of H with normalization $\alpha = \text{poly}(n)$ to be implemented in gate complexity poly(n), or alternatively, the ability to perform black-box Hamiltonian simulation for time t at gate cost $t \cdot \text{poly}(n)$.[2] In this case, an overall gate complexity for a rapidly mixing system (i.e., $t_{\text{mix}} \leq \text{poly}(n)$) can be $\text{poly}(n, \beta, 1/\epsilon)$; see [260, Table I] for a catalog of existing constructions. A recent construction [259] improved the asymptotic cost per unit Lindbladian evolution time to $\widetilde{\mathcal{O}}(\beta)$ Hamiltonian simulation time assuming black-box access to Hamiltonian simulation; for lattice Hamiltonians, this further simplifies as one merely needs to simulate lattice patches of diameter $\widetilde{\mathcal{O}}(\beta)$. To put together a practically relevant end-to-end resource estimate, one needs to design better algorithms (see, e.g., [351] for a single-ancilla variant for ground states) to reduce the per unit time cost, as well as to estimate the mixing time (e.g., by exact diagonalization of the map for small system sizes). Of course, if Gibbs sampling is employed as a heuristic (as in many classical applications of Monte Carlo methods), the cost will be empirical.

Caveats

On the one hand, the superpolynomial $\mathcal{O}(\sqrt{2^n})$ complexity for Gibbs sampling that appears explicitly in Eq. (15.1) is necessary in general (for sufficiently large β it allows one to solve NP-hard or even QMA-hard problems in the general case). On the other hand, most physical Hamiltonians (if they appear to

[2] In the model of black-box Hamiltonian simulation, one can apply the unitary $U_i = \mathrm{e}^{\mathrm{i}Ht_i}$ for user-specified choices of t_i, and the goal is to minimize the cumulative total $\sum_i t_i$ over the course of the algorithm. In this model, it is sensible for t and β to have the same units (interpreted as time), as βH and iHt are both unitless.

thermalize in nature) should be simulable without exponential hidden prefactors. The Monte Carlo–style approach to Gibbs sampling attempts to mimic nature more closely than the other algorithms with guaranteed complexities mentioned above; hence, it looks more promising for obtaining polynomial runtimes, but this must be verified through system-specific analysis or hardware demonstrations.

Finally, if the Hamiltonian comes from classical problems (such as solving semidefinite programs), loading the instance may have exponential cost ($e^{\Omega(n)}$), which in the above presentation is hidden in the assumption of a block-encoding of classical data. Additionally, it is unclear whether Hamiltonians arising from classical data—which would likely lack the local interaction structure that one sees in chemical and physical systems—should be expected to "thermalize" quickly (i.e., whether Monte Carlo–style algorithms converge in a small number of iterations).

Example use cases

- Multiplicative weights update (MWU) method and conic programming: Gibbs sampling is the main source of quantum speedup in the MWU method, which is used to solve semidefinite programs and other conic programs [181, 182, 48, 45, 46]. Existing analyses in this direction have employed Gibbs samplers with a guaranteed quadratic (but no larger) speedup, rather than the more heuristic and recent Monte Carlo–style algorithms.
- Quantum chemistry: An important step of estimating the ground state energy of electronic structure Hamiltonians is generating an ansatz state that has a large overlap with the ground state. This might be done via Gibbs sampling at sufficiently low temperatures; the overlap with the ground state is $e^{-\beta E_0}/\mathcal{Z}$, which can be large when $1/\beta$ is sufficiently small compared to the spectral gap between the ground and excited space of the Hamiltonian.
- Condensed matter physics: Similar to quantum chemistry, Gibbs sampling provides a method for producing ansatz states for ground state energy calculations, which often capture the relevant physics. However, condensed matter physicists are also interested in material properties at finite temperatures so that the Gibbs state itself might equally be of interest.
- Computing partition functions: One of the early references to develop quantum Gibbs samplers [838] applied it to the problem of estimating the partition function $\mathcal{Z}$ up to small relative error. The partition function contains all the relevant thermodynamic information of the system.
- Combinatorial optimization: Many combinatorial optimization problems can be viewed as finding the ground state of classical Hamiltonians (i.e.,

Hamiltonians that are diagonal in the computational basis), or finding a low-energy state that achieves a high approximation ratio with the ground state. Classical Monte Carlo algorithms are a key technique in this area, and quantization of these methods can sample the same classical thermal distribution with a quadratic speedup in the mixing time [944]. The full power of Gibbs sampling for general nondiagonal Hamiltonians could be useful in situations where one adds a noncommuting transverse field to the classical Hamiltonian and wishes to prepare a low-energy state, such as in quantum annealing or for training quantum Boltzmann machines [30].

Further reading

Gibbs sampling has been studied in several specific cases. For example, [152] studied Gibbs sampling of local Hamiltonians in 1D. Moreover, [603] studied *commuting* spatially local Hamiltonians and showed conditions under which they thermalize in polynomial time, suggesting efficient Gibbs sampling via Monte Carlo–style methods. These conditions hold for any 1D system at any temperature, and in any higher-spatial dimension above a certain threshold temperature.