Simulated quantum annealing can be exponentially faster than classical simulated annealing

Elizabeth Crosson

California Institute of Technology

Joint work with A. Harrow

FOCS’ 2016, arXiv: 1601.03030
Optimization and Physics

- **Combinatorial Optimization**: minimize $f : \{0, 1\}^n \rightarrow \mathbb{R}$

- **Physical inspiration**: take advantage of dynamics that drive physical systems to low energy states.

- **Classical simulated annealing (SA)**: MCMC algorithm simulates thermal cooling by attempting to sample $\pi(x) = e^{-\beta f(x)}/Z$ for $\beta_0 = 0 \leq \beta_1 \leq \ldots \leq \beta_{\text{final}}$.

- Can quantum dynamics inspire faster optimization methods?
Do quantum effects help annealing?

▶ **Intuition:** quantum dynamics allow for “tunneling” through high barriers in the energy landscape.

▶ **Theory:** QA can be exponentially faster than classical SA for particular cost functions with tall narrow energy barriers.

(FGG ‘02, Reichardt ‘04, Muthukrishnan et al. ‘15, Kong and EC ‘15, Jiang et al. ‘15, Brady and van Dam ‘16).
Adiabatic optimization and quantum annealing

- Minimize a cost function $f : \{0, 1\}^n \to \mathbb{R}$ by sampling the ground state of an $n$-qubit Hamiltonian,

\[
H_p = \sum_{z \in \{0, 1\}^n} f(z) |z\rangle\langle z|
\]

- Initialize the qubits in the ground state of a uniform transverse field $H_B = -\sum_{i=1}^n \sigma_i^x$ and interpolate from $H_B$ to $H_p$,

\[
H(s) = (1 - s)H_B + s H_p \quad , \quad 0 \leq s \leq 1
\]

- **Adiabatic optimization:** If $\Delta = \min_s \text{gap}(H(s))$ is the minimum spectral gap of $H(s)$, then time $\text{poly}(n, \Delta^{-1})$ suffices to prepare the ground state of $H_p$.

- **Quantum annealing:** includes more realistic effects e.g. being in a low temperature Gibbs state $\rho(s) = e^{-\beta H(s)}/Z(s)$. 

Exponential separation between QA and SA

- **Spike cost function**: bit-symmetric cost function with a large energy barrier that creates a local minimum.

\[
f(w) = \begin{cases} 
|w| + n^a & n/4 - n^b/2 \leq |w| \leq n/4 + n^b/2 \\
|w| & \text{otherwise}
\end{cases}
\]

- Takes time \(2^{\Omega(n^a)}\) to solve with SA, but QA succeeds in \(O(n)\) time when \(a + b < 1/2\).
QA analysis of the spike cost function

\[ C(w) = \begin{cases} 
|w| + n^a & n/4 - n^b/2 \leq |w| \leq n/4 + n^b/2 \\
|w| & \text{o.w.}
\end{cases} \]

- QA gap \( \Delta \) is constant when \( a + b < 1/2 \).
  - Spikeless Hamiltonian \( \tilde{H} \) has \( \tilde{\Delta} = \Omega(1) \).
  - \( \tilde{\psi}_0(k)^2 \) is a binomial distribution on \( k \in \{0, \ldots, n\} \), \( \implies \tilde{\psi}_0(k)^2 = \mathcal{O}(n^{-1/2}) \).
  - Spike term does not decrease the energy eigenvalues, \( E_i \geq \tilde{E}_i \).
  - Using \( \tilde{\psi}_0 \) as a variational wave function upper bounds the ground state energy by
    \[ \tilde{E}_0 + \mathcal{O}(n^a n^b n^{-1/2}) \]
    \( \implies \Delta = \Omega(1) \) when \( a + b < 1/2 \).
  - QA doesn’t even “feel” a modest sized spike at all!
However, traditional QA Hamiltonians are “stoquastic”

- **Stoquastic Hamiltonians** have real and non-positive matrix entries and can be scaled to substochastic matrices,

\[ H_{z,z'} \leq 0 \quad \text{for all} \quad z, z' \in \{0, 1\}^n. \]

- Thermal state \( \rho = e^{-\beta H} / tr(e^{-\beta H}) \) is a non-negative matrix, and so \( |\psi_0\rangle \langle \psi_0| = \lim_{\beta \to \infty} \rho \) has non-negative amplitudes.

- Adiabatic computation with *frustration-free* stoquastic \( H \) can be classically simulated in poly time (Bravyi and Terhal, 2008)

- Stoquastic LH problem is in AM (Bravyi et al., 2006)

- **Open question:** Can stoquastic adiabatic computation be classically simulated in time \( \text{poly}(n, \Delta^{-1}) \)?
Path integrals and importance sampling

- Express \( Z = \text{tr} \left( e^{-\beta H} \right) \) as a discrete path-integral:

\[
Z = \sum_{z \in \{0,1\}^n} \langle z | \left( e^{-\frac{\beta H}{L}} \right)^L | z \rangle = \sum \prod_{i=1}^{L} \langle z_i | e^{-\frac{\beta H}{L}} | z_{i+1} \rangle
\]

- Since \( H \) is stoquastic the terms of the sum are all positive, leading to a probability distribution on the space of paths,

\[
\pi(z_1, ..., z_L) = \frac{1}{Z} \prod_{i=1}^{L} \langle z_i | e^{-\frac{\beta H}{L}} | z_{i+1} \rangle
\]

- Stoquastic systems are common in nature (e.g. spinless particles), and in physics these thermal path integrals are often viewed as being in “Euclidean” or “imaginary” time.

- Evidently, some paths are more important!
Simulated Quantum Annealing

SQA discretizes the adiabatic path and samples $\pi$ at each point using a Markov chain. The samples can be used for Monte Carlo estimation of physical observables.

$$\pi(z_1, ..., z_L) = \frac{1}{Z} \prod_{i=1}^{L} \langle z_i | e^{-\beta H} | z_{i+1} \rangle , \quad z_i \in \{0, 1\}^n$$

$\pi$ penalizes configurations for the proportion of “time” they spend on bit strings with high $f$, and also for the total number of “jumps” along the path.
SQA for the spike converges in poly time

**Our result:** SQA equilibrates in polynomial time and finds the minimum of the spike cost function whenever $a + b < 1/2$.

- $\tilde{O}(n^{17})$ with single-site Metropolis updates

- $\tilde{O}(n^7)$ with worldline heat-bath updates
Visualizing SQA for the spike system

\[ s \approx 0 \]

\[ s \approx 0.3 \]

\[ s \approx 0.36 \]

\[ s \approx 0.5 \]
Markov chains and Mixing times

- **Markov chain**: sample from $\pi$ using a random walk on $\Omega$, with transition probabilities satisfying detailed balance,

$$\pi(x) P(x, y) = P(y, x) \pi(y).$$

- **Mixing time**: number of steps of the random walk which are needed to be close in trace distance to $\pi$,

$$\tau(\epsilon) = \max_x \min_t \{ t : \|P^t(x, \cdot) - \pi\|_1 \leq \epsilon \ \text{for all} \ t' \geq t \}$$

- **Transition matrix** $P = \sum P(x, y) |x\rangle\langle y|$ has largest eigenvalue 1, corresponding to $\langle \pi | = \sum \pi(x) |x\rangle$, and the gap $\lambda$ to the next largest eigenvalue determines the mixing time,

$$\tau(\epsilon) \leq \lambda^{-1} \log \left( \frac{1}{\epsilon \ \pi_{\min}} \right)$$
Conductance and Bottlenecks

- **Visualize** $(\pi, P, \Omega)$ as a weighted graph
  - Vertices are the states $x \in \Omega$
  - Connect vertices $x, y \in \Omega$ with an edge iff $P(x, y) > 0$

- The **Conductance** of a set $S \subseteq \Omega$ with $\pi(S) \leq 1/2$ is the probability of leaving $S$ over the probability of being in $S$,
  \[
  \Phi(S) := \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x)P(x, y)
  \]

- **Cheeger’s inequality** relates the minimum conductance $\Phi_{\text{min}} := \min_S \Phi(S)$ to the spectral gap, $\lambda = \text{gap}(P)$,
  \[
  \frac{\Phi_{\text{min}}^2}{2} \leq \lambda \leq 2\Phi_{\text{min}}
  \]

- Rapid mixing iff state space graph has no “bottlenecks”!
Canonical paths: duality between cuts and flows

- **Design paths** $\gamma_{xy}$ for all $x, y \in \Omega$ that carry traffic $\pi(x)\pi(y)$ and don’t overload any edge capacities $\pi(z)P(z, z')$.

- **Congestion** $R$ is the worst-case ratio of the total traffic of paths passing through an edge to the capacity of that edge,

$$R = \max_{(z,z')} \frac{1}{\pi(z)P(z, z')} \sum_{\gamma_{xy} \ni (z,z')} \pi(x)\pi(y)|\gamma_{xy}|$$

$$\tau_{\text{mix}} = O(R \log \pi^{-1}_{\text{min}})$$
Overview of the SQA analysis

- Compare the SQA Markov chain with and w/o the spike term.

- For “worst-case” configurations the spike distribution $\pi$ may be very different from the spikeless distribution $\tilde{\pi}$.

- Use expectation values of the quantum system to see that for ”typical configurations” $\pi \approx \tilde{\pi}$ when $a + b < 1/2$.

- Use canonical paths for the spikeless system to construct canonical paths for the spike system within the typical subset, and show that leaks outside of this subset are rare.

- Quasi-equilibration of SQA within a subset of the state space, adiabatic path guarantees warm starts in this subset.
Most-paths comparison method

- **Idea:** $z \in \Omega$ which do not spend too much proper time on the spike will have $\pi(z)$ close to the spikeless distribution $\tilde{\pi}(z)$.

- Using “most” of the canonical paths for the spikeless system we construct paths for the spike system with $R = \mathcal{O}(\tilde{R})$ inside a subset $\Omega_G$ of large measure, $\pi(\Omega_G) > 1 - 1/\text{poly}(n)$.
Congestion of the spikeless chain

- **Heat-bath worldline updates** \((\bar{x}_1, \ldots, \bar{x}_n) \rightarrow (\bar{x}'_1, \ldots, \bar{x}'_n)\).
  1. Select \(i \in \{1, \ldots, n\}\) u.a.r.
  2. Set \(\bar{x}'_j = \bar{x}_j\) for all \(j \neq i\)
  3. Sample \(\bar{x}'_i\) from \(\pi(\bar{x}'_i|\bar{x}'_1, \ldots, \bar{x}'_{i-1}, \bar{x}'_{i+1}, \ldots, \bar{x}'_n)\).

- **Congestion** is \(\tilde{R} = \mathcal{O}(n^2)\) because the paths have \(n\) steps, and each step has transition probability \(\mathcal{O}(1/n)\).

- **Implemented** in \(\tilde{O}(n^2 \beta^2)\) time (without using symmetry)
  - \(\mathcal{O}(\beta \log n)\) jumps per worldline
  - break worldline \(i\) into \(\mathcal{O}(n \beta \log n)\) segments
  - Checking all spin values in each segment takes time \(\mathcal{O}(n)\)
  - Generate \(\mathcal{O}(\beta \log n)\) new jumps in worldline \(i\) and flip a weighted coin to assign a bit value to each time segment
Total run-time of SQA for the spike with worldline updates

- \( \tilde{R} = O(n^2) \) and \( R = O(\tilde{R}) \) in \( \Omega_G \) \( \implies \) \( \tau_{\text{mix}} = O(n^2 \log \pi_{\text{min}}^{-1}) \) within \( \Omega_G \), at each step of the adiabatic path.

- \( O(\beta \log n) \) jumps per worldline implies \( \log (\pi_{\text{min}}^{-1}) = \tilde{O}(n\beta) \).

- Suffices to take \( \tilde{O}(n\beta) \) evenly spaced steps along the adiabatic path to fulfill the warm start condition.

- One step of worldline update chain takes time \( \tilde{O}(n^2\beta^2) \).

\[ \beta = n^{\frac{1}{4} - \frac{a}{2}} \implies \tilde{O}(n^7) \text{ total run-time!} \]
Summary and Conclusion

- We’ve shown that SQA can inherit some of the quantum advantages of QA and given the first proof of an exponential separation in the asymptotic performance of SA and SQA.

- Can we apply the ideas in the proof to better understand the stationary distribution and convergence properties of SQA in more general systems?

- Besides path integral SQA, there are other “quantum Monte Carlo” methods for stoquastic Hamiltonians. New ideas are needed, however, since no existing QMC method can efficiently simulate stoquastic QA in full generality.

- Thank you for your attention!