On Quantum Speedups for Nonconvex Optimization via Quantum Tunneling Walks

Quantum 7, 1030 (2023)
arXiv:2209.14501

Joint work with Yizhou Liu (Tsinghua U/MIT) and Weijie J. Su (U Penn)

Tongyang Li
Peking University

IPAM Workshop: Mathematical Aspects of Quantum Learning, October 18, 2023

Outline

1. Motivations

2. Preliminaries: Quantum walks

3. Quantum tunneling walks

4. Applications and numerical experiments

Optimization

Problem: $f\colon \mathbb{R}^n \to \mathbb{R}, \quad \min_x f(x)$

Core topic in applied mathematics, computer science, physics, etc.

Provable guarantee for solving an optimization problem?

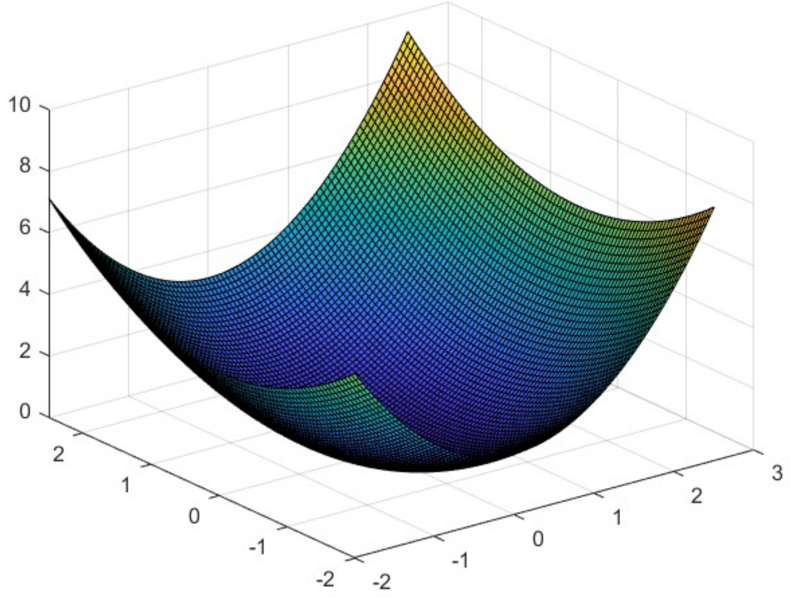
Convex optimization can be solved in polynomial time if we can query f: input x, output $f(x)$. Methods: ellipsoid method, interior point method, etc.

Nonconvex Optimization

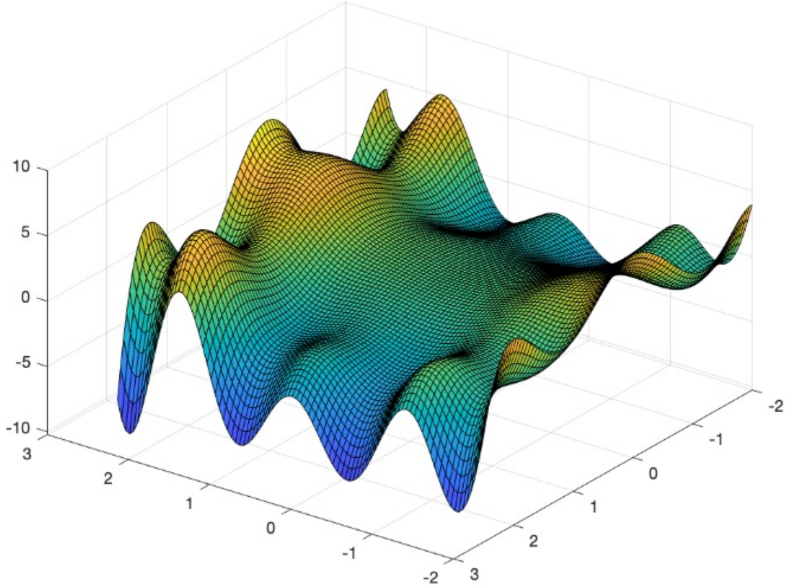
However, in practice, many functions are **nonconvex.**

For instance, in machine learning: Train an ML model ⟷ Optimize a loss function

Loss functions of neural networks: very nonconvex in general.



Unique local minimum is the global one



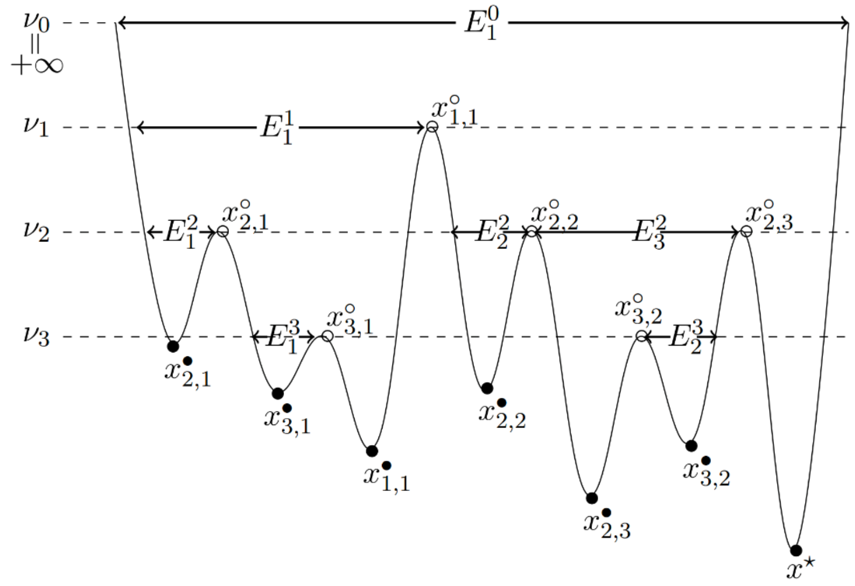
So many local minima…

Nonconvex Optimization

Stochastic gradient descent (SGD): $x_{k+1} = x_k - s\nabla f(x_k) - s\xi_k$ with learning rate s and the kth step noise ξ_k.

Working horse in training neural networks, e.g., ADAM, Adagrad, are improved versions of SGD.

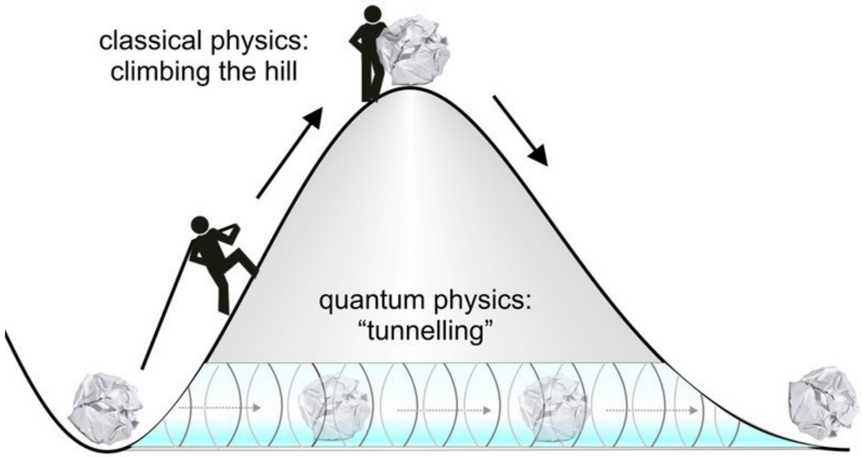
Most common issue: Exploration in the optimization landscape.



From the left side, need to jump through all the local minima and maxima in the middle to reach x^*

Quantum Tunneling

In quantum physics, nonconvex landscapes appear in quantum tunneling.



The Schrödinger equation: $i\frac{\partial}{\partial t}\Phi = \left(-\underset{\downarrow}{h^2}\Delta + f(x)\right)\Phi$

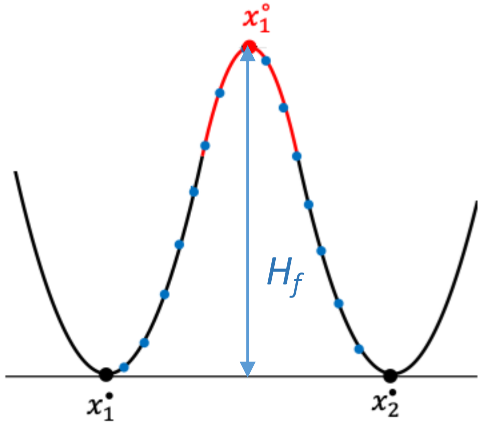
quantum learning rate

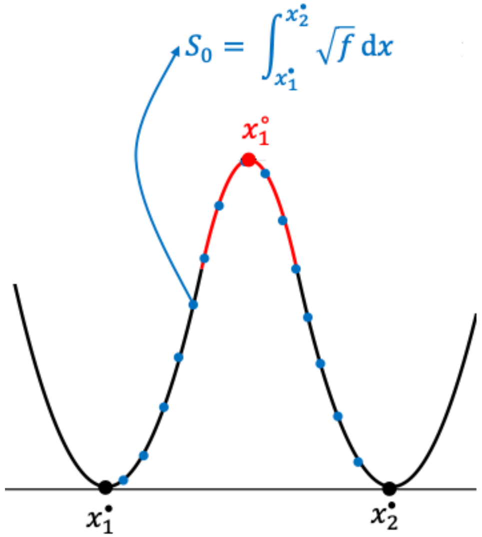
For 1-dimension potential, WKB approximation (semiclassical analysis) gives:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\Psi(x)+V(x)\Psi(x)=E\Psi(x) \Longrightarrow \boxed{\Psi(x)\approx\frac{C_+e^{\int \hbar^{-1}\sqrt{2m(V(x)-E)}\,dx}+C_-e^{-\int \hbar^{-1}\sqrt{2m(V(x)-E)}\,dx}}{\hbar^{-1/2}\sqrt[4]{2m\left(V(x)-E\right)}}.}$$

Comparisons

	Classical	**Quantum**
Dynamics (continuous-time)	$\mathrm{d}x = -\nabla f(x)\mathrm{d}t + \sqrt{s}\mathrm{d}W$ Langevin equation	$i\frac{\partial}{\partial t}\Phi = \left(-h^2\Delta + f(x)\right)\Phi$ Schrödinger equation
Convergence rate (1-dimension f)	$\exp(H_f / s)$	$\exp(S_0 / \mathrm{h})$





$S_0 = \int_{x_1^{\bullet}}^{x_2^{\bullet}} \sqrt{f}\,\mathrm{d}x$
x_1°
$x_1^{\bullet}$
$x_2^{\bullet}$

Optimization of High-Dimensional Nonconvex Functions

First: Can we efficiently simulate the Schrödinger equation with high-dimensional potential?

Assumption: Quantum evaluation oracle $O_f \left|x\right\rangle \left|z\right\rangle = \left|x\right\rangle \left|f(x) + z\right\rangle \quad \forall x \in \mathbb{R}^d, z \in \mathbb{R}.$

- It allows *coherent superpositions* of queries to *f*, a standard assumption for quantum algorithms working in real space.
- In practice, real numbers are represented digitally, but we assume the representation has *sufficiently high precision* that errors from this digital representation can be neglected.
- If *f* can be computed by a classical circuit, then the corresponding quantum oracle can be implemented by a quantum circuit of *roughly the same size*.

The cost of simulating the Schrodinger equation for time t:

$$O\left(\|f\|_{L^\infty(\Omega)} t \frac{\log(\|f\|_{L^\infty(\Omega)} t/\epsilon)}{\log\log(\|f\|_{L^\infty(\Omega)} t/\epsilon)}\right)$$

Optimization of High-Dimensional Nonconvex Functions

Question: What is the behavior of the Schrödinger equation for high-dimensional functions?

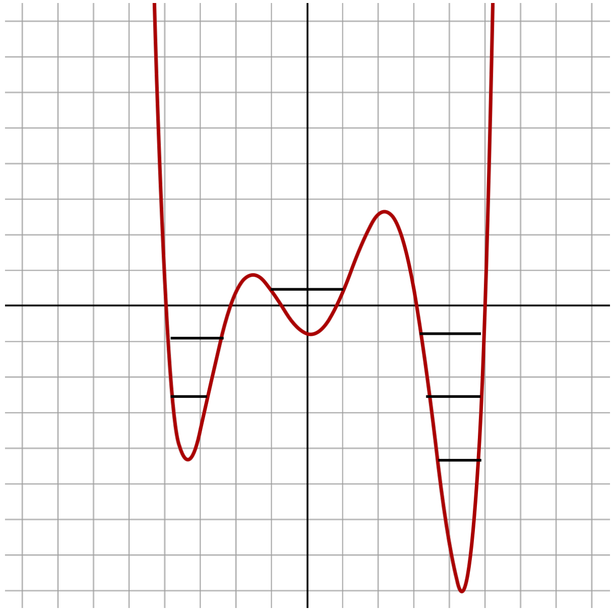
Main result (informal): For nonconvex functions whose local minima have equal values, i.e., all local minima are global, the evolution of the Schrödinger equation behaves as:

quantum tunneling + quantum walk = quantum tunneling walk

Quantum tunneling: the strength of tunneling can be described by the 1-dim WKB.

Quantum walk: Having the tunneling strength between wells, the amplitudes evolve like a continuous-time quantum walk.

Quantum Tunneling Walk

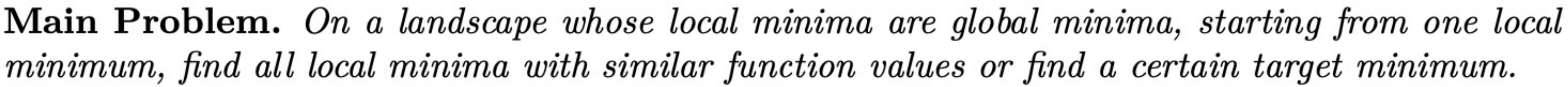


Q: Why do we require local minima to be global?

A: Otherwise, there can be nontrivial probability transiting form the ground state of one well to excited states of another well.

Q: If all local minima are global, why do we want to study this problem?

A: Optimization is only one of goals of nonconvex problems. Generalization is also important.



Main Problem. *On a landscape whose local minima are global minima, starting from one local minimum, find all local minima with similar function values or find a certain target minimum.*

Main Result

Theorem 1 (Quantum tunneling walks, informal)

On landscapes whose local minima are global minima, we have an algorithm called quantum tunneling walks (QTW) which initiates the simulation of the Schrödinger equation from the local ground state at a minimum, and measures the position at a time which is chosen uniformly from $[0, \tau]$. To solve the Main Problem we can take

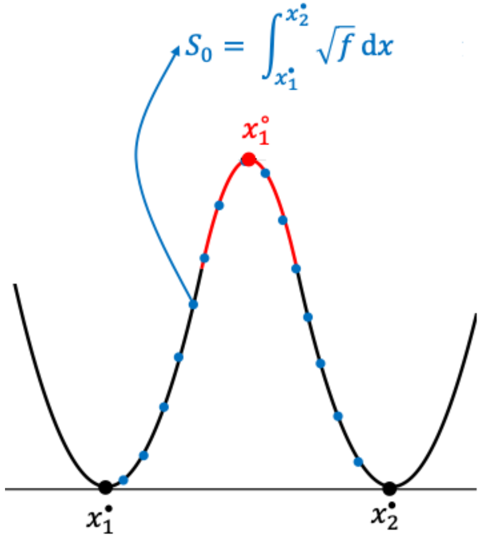
$$\tau = O(\mathrm{poly}(N)/\Delta E),$$

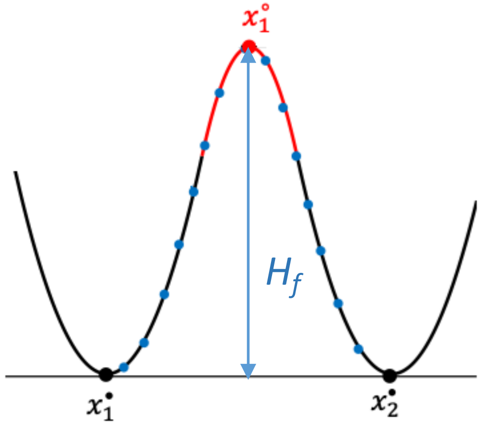
where N is the number of global minima and ΔE is the minimal spectral gap of the Hamiltonian restricted in a low-energy subspace. For sufficiently small h, we have

$$\Delta E = \sqrt{h}(b + O(h))e^{-\frac{S_0}{h}},$$

where $b, S_0 > 0$ are constants that depend only on f.

More Comparisons


$S_0 = \int_{x_1^\bullet}^{x_2^\bullet} \sqrt{f}\, \mathrm{d}x$
x_1°
$x_1^\bullet$
$x_2^\bullet$



This work

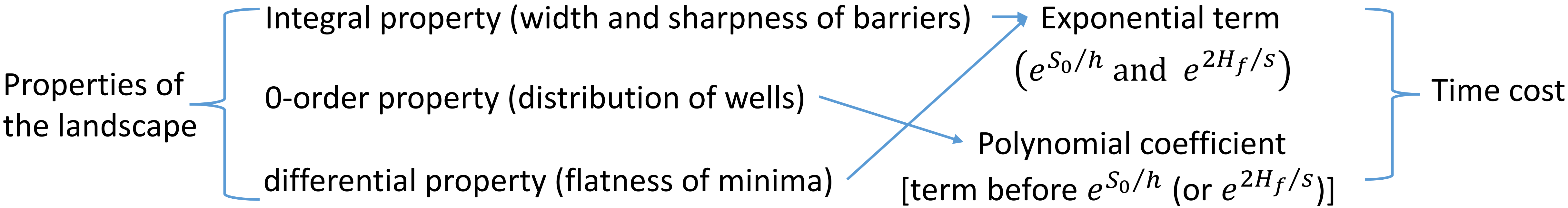
Q time cost = $O(\text{poly}(N)/\Delta E)$ $\Delta E = \sqrt{h}(b + O(h))e^{-\frac{S_0}{h}}$

vs.

[Shi, Su, and Jordan, arXiv:2004.06977]

C time cost = $O(1/\lambda_s)$ $\lambda_s = (a + o(s))e^{-\frac{2H_f}{s}}$

Main Message



Properties of the landscape

Integral property (width and sharpness of barriers)

0-order property (distribution of wells)

differential property (flatness of minima)

Exponential term ($e^{S_0/h}$ and $e^{2H_f/s}$)

Polynomial coefficient [term before $e^{S_0/h}$ (or $e^{2H_f/s}$)]

Time cost

Outline

1. Motivations

2. Preliminaries: Quantum walks

3. Quantum tunneling walks

4. Applications and numerical experiments

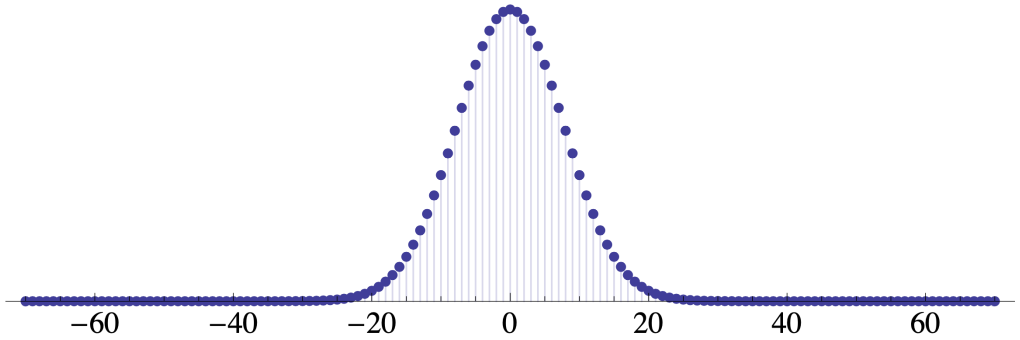
Quantum Walks

Quantum analog of a random walk on a graph, but replace probabilities by amplitudes.

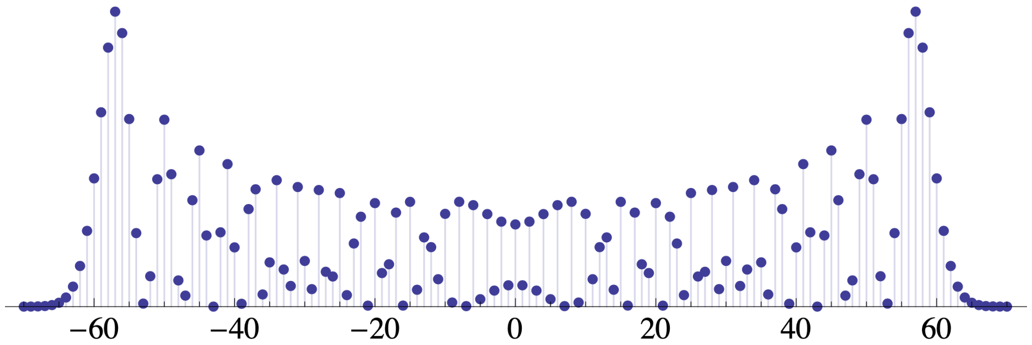
Interference can produce radically different behavior!

Ex. One-dimensional walks

Classical

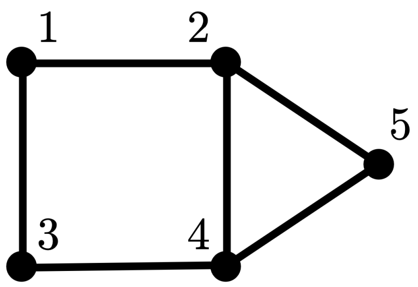


Quantum



Quantum Walks on Graphs

Graph G:



Laplacian L of G: $$L = \begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 1 & -3 & 0 & 1 & 1 \\ 1 & 0 & -2 & 1 & 0 \\ 0 & 1 & 1 & -3 & 1 \\ 0 & 1 & 0 & 1 & -2 \end{pmatrix}$$

Classical random walk

State: Probability $P_v(t)$ of being at vertex v at time t

Dynamics: $\frac{\mathrm{d}}{\mathrm{d}t}\vec{p} = L\vec{p}$

Quantum walk

State: Amplitude $a_v(t)$ of being at vertex v at time t

Dynamics: $i\frac{\mathrm{d}}{\mathrm{d}t}\vec{a} = L\vec{a}$

Mixing Time

$$L = \begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 1 & -3 & 0 & 1 & 1 \\ 1 & 0 & -2 & 1 & 0 \\ 0 & 1 & 1 & -3 & 1 \\ 0 & 1 & 0 & 1 & -2 \end{pmatrix}$$

The Laplacian L is negative semidefinite.

0 is an eigenvalue with the uniform vector *u*=(1/n,…,1/n) being its eigenvector.

Furthermore, if G is connected, then 0 has single multiplicity, i.e., all other eigenvectors of G has eigenvalue ≠ 0. Random walk mixes to *u*:

$$\begin{aligned} p(t) &= e^{Lt}p(0) \\ &= \left(|V|uu^T + \sum_{\lambda\neq 0} e^{\lambda t} v_\lambda v_\lambda^T\right) p(0) \\ &= \langle |V|u, p(0)\rangle u + \sum_{\lambda\neq 0} e^{\lambda t}\langle v_\lambda, p(0)\rangle v_\lambda \\ &= u + \sum_{\lambda\neq 0} e^{\lambda t}\langle v_\lambda, p(0)\rangle v_\lambda \end{aligned}$$

Convergence rate: O(log(1/ε)/Δ), where Δ is the spectral gap of G.

Mixing Time of Quantum Walks

In quantum computing, the story is totally different.

Fact: The Schrödinger equation $i\frac{\mathrm{d}}{\mathrm{d}t}\vec{a} = H\vec{a} \Rightarrow \vec{a}(t) = e^{-iHt}\vec{a}(0)$

gives a unitary dynamics that keeps rotating and does not converge.

Solution: Define a notion of limiting distribution by taking the end time t uniformly in [0,T].

$$\begin{aligned}
p_{a\to b}(T) &= \frac{1}{T}\int_0^T |\langle b|e^{-iHt}|a\rangle|^2 \mathrm{d}t \\
&= \sum_{\lambda,\lambda'} \langle b|\lambda\rangle\langle\lambda|a\rangle\langle a|\lambda'\rangle\langle\lambda'|b\rangle \frac{1}{T}\int_0^T e^{-i(\lambda-\lambda')t}\mathrm{d}t \\
&= \sum_{\lambda} |\langle a|\lambda\rangle\langle b|\lambda\rangle|^2 + \sum_{\lambda\neq\lambda'} \langle b|\lambda\rangle\langle\lambda|a\rangle\langle a|\lambda'\rangle\langle\lambda'|b\rangle \frac{1-e^{-i(\lambda-\lambda')T}}{i(\lambda-\lambda')T}
\end{aligned}$$

Mixing Time of Quantum Walks

$$\begin{aligned}
p_{a\to b}(T) &= \frac{1}{T}\int_0^T |\langle b|e^{-iHt}|a\rangle|^2 \mathrm{d}t \\
&= \sum_{\lambda,\lambda'} \langle b|\lambda\rangle\langle\lambda|a\rangle\langle a|\lambda'\rangle\langle\lambda'|b\rangle \frac{1}{T}\int_0^T e^{-i(\lambda-\lambda')t}\mathrm{d}t \\
&= \sum_{\lambda} |\langle a|\lambda\rangle\langle b|\lambda\rangle|^2 + \sum_{\lambda\neq\lambda'} \langle b|\lambda\rangle\langle\lambda|a\rangle\langle a|\lambda'\rangle\langle\lambda'|b\rangle \frac{1-e^{-i(\lambda-\lambda')T}}{i(\lambda-\lambda')T}
\end{aligned}$$

If Δ is the smallest gap between any pair of eigenvalues of A (Δ>0 implies that A is non-degenerate), then: $p_{a\to b}(\infty) := \sum_{\lambda} |\langle a|\lambda\rangle\langle b|\lambda\rangle|^2.$

Furthermore:

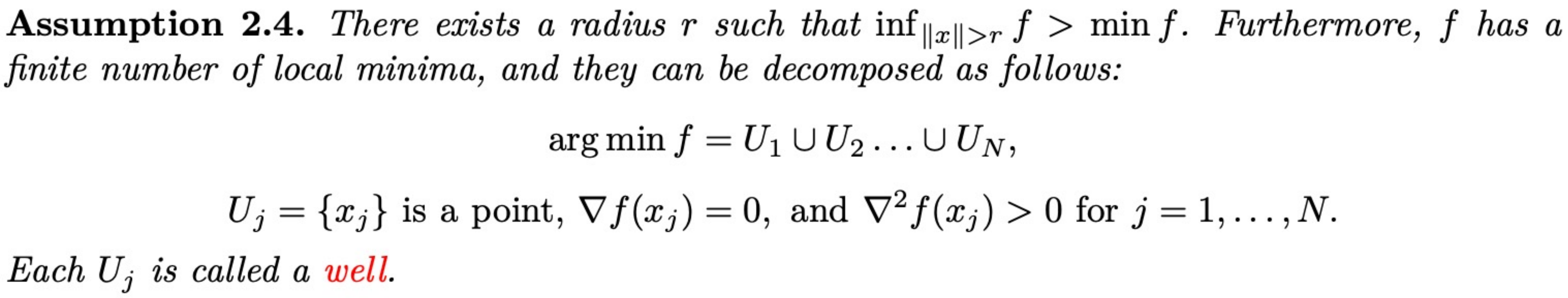
$$\begin{aligned}
|p_{a\to b}(T) - p_{a\to b}(\infty)| &\leq \sum_{\lambda\neq\lambda'} |\langle b|\lambda\rangle\langle\lambda|a\rangle\langle a|\lambda'\rangle\langle\lambda'|b\rangle| \cdot \left|\frac{1-e^{-i(\lambda-\lambda')T}}{i(\lambda-\lambda')T}\right| \\
&\leq \frac{2}{T\Delta}\sum_{\lambda,\lambda'} |\langle b|\lambda\rangle\langle\lambda|a\rangle| \cdot |\langle a|\lambda'\rangle\langle\lambda'|b\rangle| \\
&= \frac{2}{T\Delta}|\langle b|a\rangle|^2 \leq \frac{2}{T\Delta}.
\end{aligned}$$

Outline

1. Motivations
2. Preliminaries: Quantum walks
3. Quantum tunneling walks
4. Applications and numerical experiments

Assumptions on the Nonconvex Function

The assumption that “all local minima are global” is formally stated as follows:



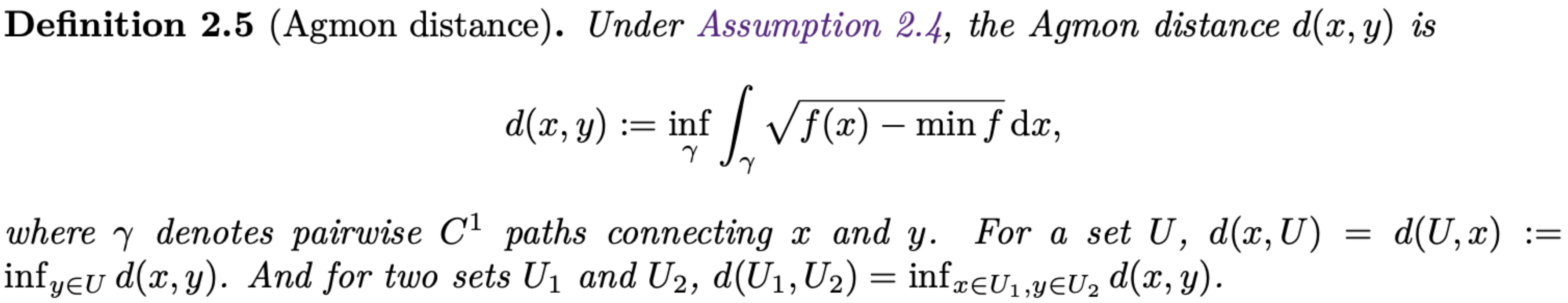
Assumption 2.4. *There exists a radius r such that* $\inf_{\|x\|>r} f > \min f$. *Furthermore, f has a finite number of local minima, and they can be decomposed as follows:*

$$\arg\min f = U_1 \cup U_2 \ldots \cup U_N,$$

$$U_j = \{x_j\} \text{ is a point, } \nabla f(x_j) = 0, \text{ and } \nabla^2 f(x_j) > 0 \text{ for } j = 1, \ldots, N.$$

Each U_j is called a well.

We can then further define the distances as follows:

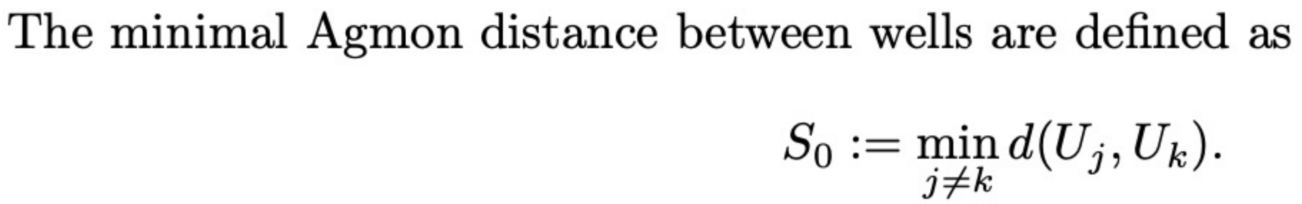


Definition 2.5 (Agmon distance)**.** *Under Assumption 2.4, the Agmon distance $d(x, y)$ is*

$$d(x, y) := \inf_{\gamma} \int_{\gamma} \sqrt{f(x) - \min f}\, \mathrm{d}x,$$

where γ denotes pairwise C^1 paths connecting x and y. For a set U, $d(x, U) = d(U, x) :=$ $\inf_{y \in U} d(x, y)$. *And for two sets U_1 and U_2,* $d(U_1, U_2) = \inf_{x \in U_1, y \in U_2} d(x, y)$.

Assumptions on the Nonconvex Function



The minimal Agmon distance between wells are defined as

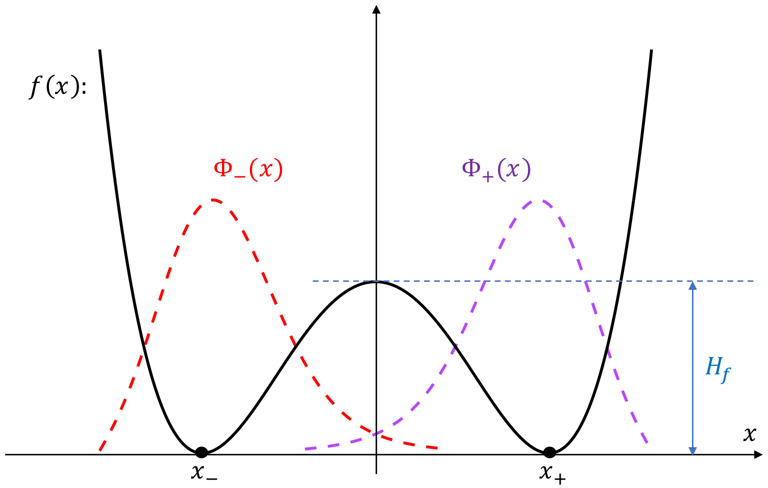
$$S_0 := \min_{j \neq k} d(U_j, U_k).$$

Next, we want to further look at the local landscape near a well.
For a small enough η, we consider $B(U_j, \eta) = \{x \in M | d(x, U_j) \leq \eta\}$

For each well j, we can hence define its local ground state $|e_j\rangle$ as its eigenstate with the minimum eigenvalue.

$$i\frac{\partial}{\partial t}\Phi = \left(-h^2\Delta + f(x)\right)\Phi$$

Fact 1: For sufficiently small h in the Schrödinger equation, the local ground states $|e_1\rangle$, $|e_2\rangle$, … , $|e_N\rangle$ almost localize near the wells U_1, U_2, … , U_N: Exponential decay with respect to the Agmon distance to its corresponding well.



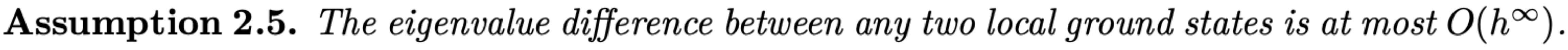
Assumptions on the Nonconvex Function

Fact 2: The space $\mathcal{F}$ spanned by $\{|e_1\rangle, |e_2\rangle, \ldots, |e_N\rangle\}$ is a low-energy invariant subspace of the Hamiltonian $H=-h^2\Delta+f(x)$. In other words, in the low-energy space $\mathcal{F}$, the particle walks between wells by quantum tunneling.

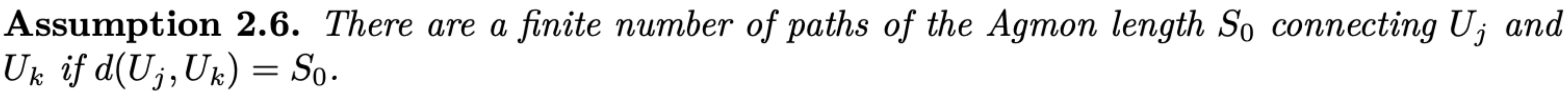
Fact 3: The Hamiltonian restricted in $\mathcal{F}$, i.e., $H_{|\mathcal{F}}$, determines the strength of the quantum tunneling effect and is called the *interaction matrix*.

Two further assumptions:

dominated by h^n
for any $n>0$



Assumption 2.5. *The eigenvalue difference between any two local ground states is at most $O(h^\infty)$.*

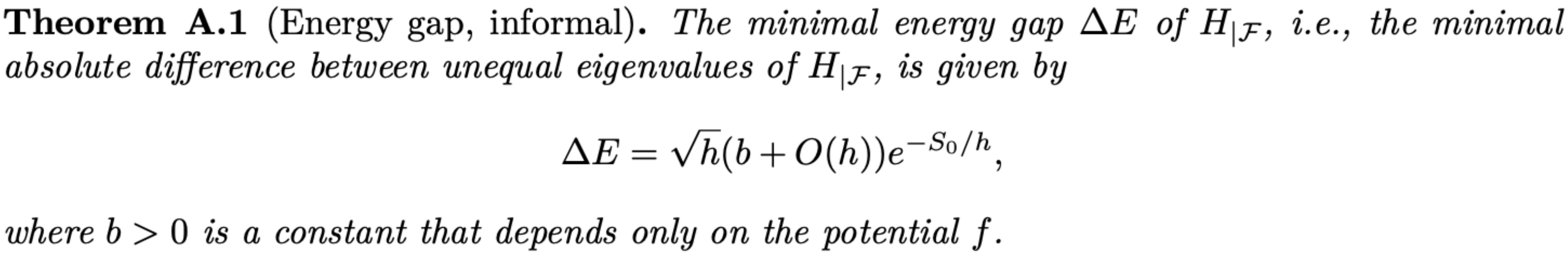


Assumption 2.6. *There are a finite number of paths of the Agmon length S_0 connecting U_j and U_k if $d(U_j, U_k) = S_0$.*

Assumptions on the Nonconvex Function

Based on our assumptions + existing literature in functional analysis, we use semiclassical analysis (WKB approximation) to analyze the tunneling effect in $H_{|\mathcal{F}}$.

Our main result for spectrum properties of the Schrodinger Hamiltonian $H=-h^2\Delta+f(x)$:



Theorem A.1 (Energy gap, informal). *The minimal energy gap ΔE of $H_{|\mathcal{F}}$, i.e., the minimal absolute difference between unequal eigenvalues of $H_{|\mathcal{F}}$, is given by*

$$\Delta E = \sqrt{h}(b + O(h))e^{-S_0/h},$$

where $b > 0$ is a constant that depends only on the potential f.

This can be naturally applied to analyzing the mixing time of the quantum tunneling walk!

Mixing Time of Quantum Tunneling Walks

Let the spectral decomposition of $H_{|\mathcal{F}}$ to be $H_{|\mathcal{F}} = \sum_{k=1}^{N} E_k \left|E_k\right\rangle \left\langle E_k\right|$.

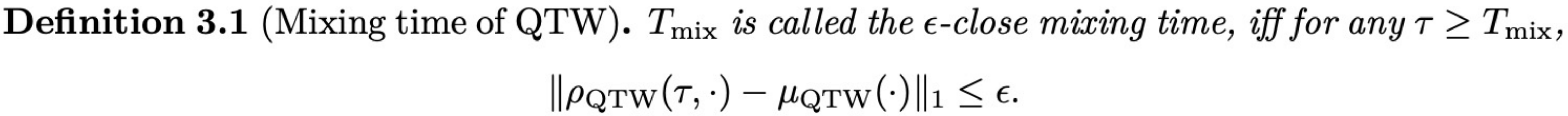
Choosing t uniformly in $[0, \tau]$, the probability density of finding the walker at x is

$$\begin{aligned}\rho_{\mathrm{QTW}}(\tau, x) :=& \frac{1}{\tau} \int_0^\tau \mathrm{d}t |\langle x| e^{-iH_{|\mathcal{F}} t} \left|\Phi(0)\right\rangle|^2 \\ =& \sum_{E_k = E_{k'}} \langle x|E_k\rangle\langle E_k|\Phi(0)\rangle\langle\Phi(0)|E_{k'}\rangle\langle E_{k'}|x\rangle \\ &+ \sum_{E_k \neq E_{k'}} \frac{1 - e^{-i(E_k - E_{k'})\tau}}{i(E_k - E_{k'})\tau} \langle x|E_k\rangle\langle E_k|\Phi(0)\rangle\langle\Phi(0)|E_{k'}\rangle\langle E_{k'}|x\rangle.\end{aligned}$$

The time-averaged probability density leads to a limiting distribution when $\tau \to \infty$:

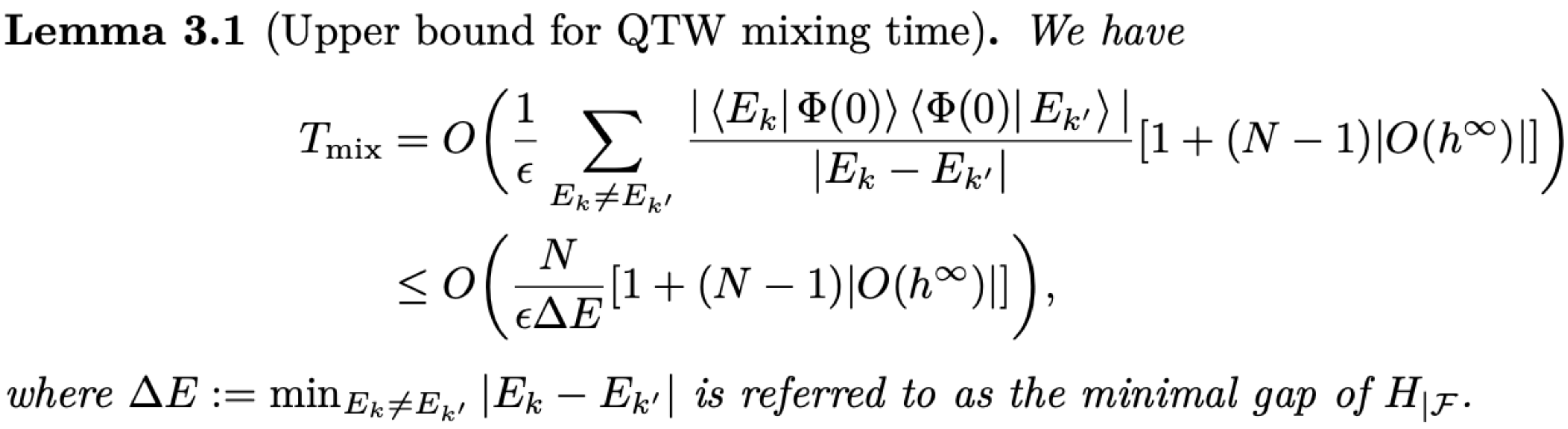
$$\mu_{\mathrm{QTW}} := \sum_{E_k = E_{k'}} \langle x|E_k\rangle\langle E_k|\Phi(0)\rangle\langle\Phi(0)|E_{k'}\rangle\langle E_{k'}|x\rangle.$$

Mixing Time of Quantum Tunneling Walks



Definition 3.1 (Mixing time of QTW). *T_{mix} is called the ϵ-close mixing time, iff for any $\tau \geq T_{\text{mix}}$,*

$$\|\rho_{\text{QTW}}(\tau, \cdot) - \mu_{\text{QTW}}(\cdot)\|_1 \leq \epsilon.$$



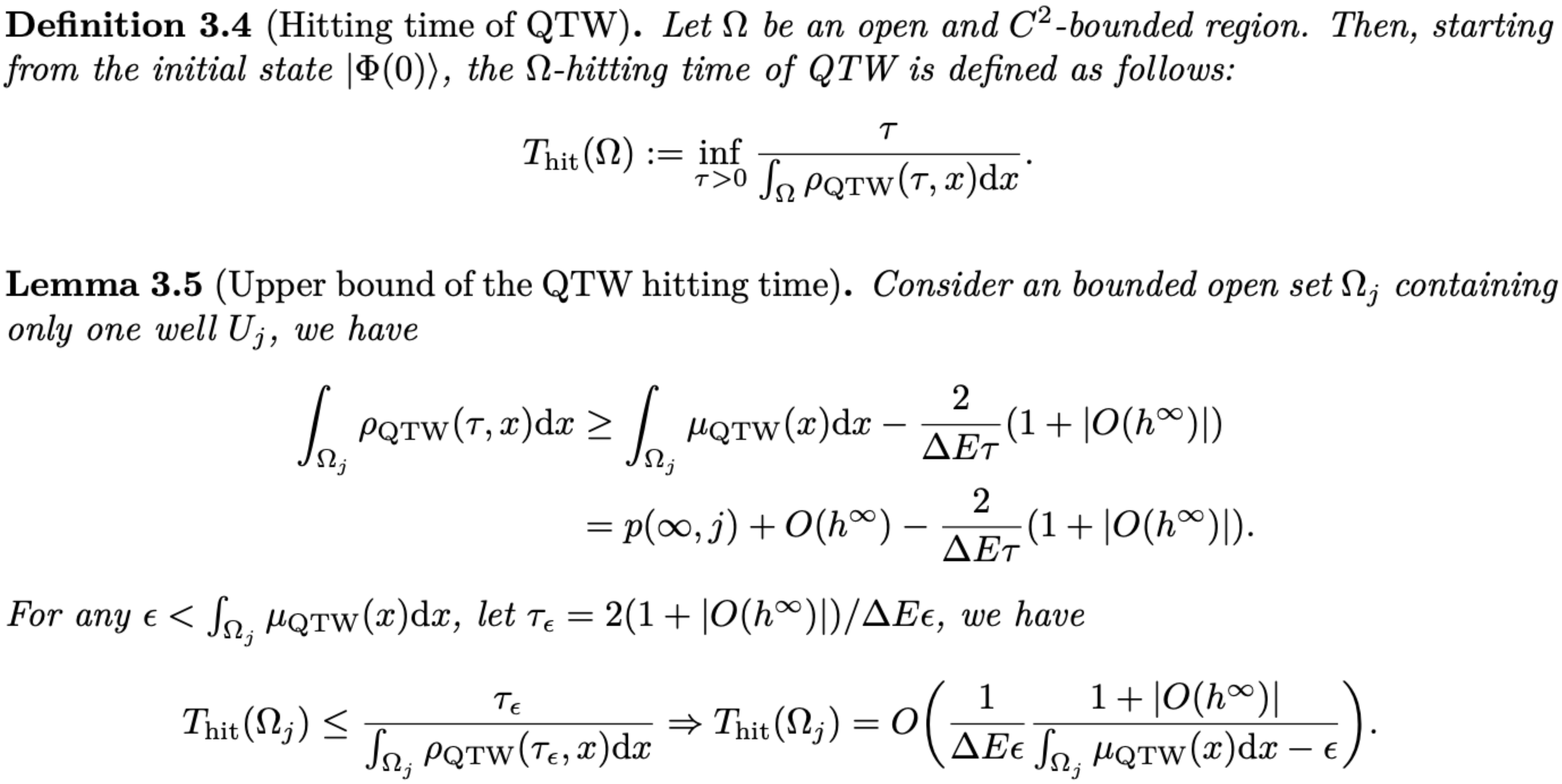
Lemma 3.1 (Upper bound for QTW mixing time). *We have*

$$\begin{aligned} T_{\text{mix}} &= O\left(\frac{1}{\epsilon} \sum_{E_k \neq E_{k'}} \frac{|\langle E_k|\Phi(0)\rangle \langle\Phi(0)|E_{k'}\rangle|}{|E_k - E_{k'}|}[1 + (N-1)|O(h^\infty)|]\right) \\ &\leq O\left(\frac{N}{\epsilon \Delta E}[1 + (N-1)|O(h^\infty)|]\right), \end{aligned}$$

where $\Delta E := \min_{E_k \neq E_{k'}} |E_k - E_{k'}|$ is referred to as the minimal gap of $H_{|\mathcal{F}}$.

As quantum walk mixes, it can also hit a specific set.

Hitting Time of Quantum Tunneling Walks



Definition 3.4 (Hitting time of QTW). *Let Ω be an open and C^2-bounded region. Then, starting from the initial state $|\Phi(0)\rangle$, the Ω-hitting time of QTW is defined as follows:*

$$T_{\mathrm{hit}}(\Omega) := \inf_{\tau > 0} \frac{\tau}{\int_\Omega \rho_{\mathrm{QTW}}(\tau, x)\mathrm{d}x}.$$

Lemma 3.5 (Upper bound of the QTW hitting time). *Consider an bounded open set Ω_j containing only one well U_j, we have*

$$\begin{aligned}\int_{\Omega_j} \rho_{\mathrm{QTW}}(\tau, x)\mathrm{d}x &\geq \int_{\Omega_j} \mu_{\mathrm{QTW}}(x)\mathrm{d}x - \frac{2}{\Delta E \tau}(1 + |O(h^\infty)|) \\ &= p(\infty, j) + O(h^\infty) - \frac{2}{\Delta E \tau}(1 + |O(h^\infty)|).\end{aligned}$$

For any $\epsilon < \int_{\Omega_j} \mu_{\mathrm{QTW}}(x)\mathrm{d}x$, let $\tau_\epsilon = 2(1 + |O(h^\infty)|)/\Delta E \epsilon$, we have

$$T_{\mathrm{hit}}(\Omega_j) \leq \frac{\tau_\epsilon}{\int_{\Omega_j} \rho_{\mathrm{QTW}}(\tau_\epsilon, x)\mathrm{d}x} \Rightarrow T_{\mathrm{hit}}(\Omega_j) = O\left(\frac{1}{\Delta E \epsilon} \frac{1 + |O(h^\infty)|}{\int_{\Omega_j} \mu_{\mathrm{QTW}}(x)\mathrm{d}x - \epsilon}\right).$$

Outline

1. Motivations

2. Preliminaries: Quantum walks

3. Quantum tunneling walks

4. Applications and numerical experiments

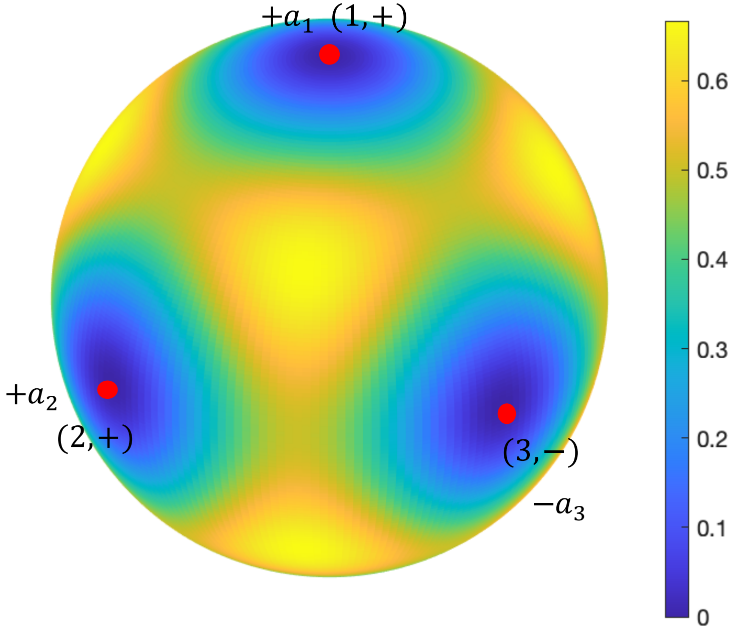
Applications

Two main questions:

1. Are their natural high-dim nonconvex optimization problems falling into our theory?
2. Can we achieve significant quantum speedups for certain problems?

Short answer: Yes for both questions, though not very satisfactory.

Application: Orthogonal Tensor Decomposition



Central problem in learning latent variable models

$$T = \sum_{i=1}^{d} a_j^{\otimes 4}$$

$$f(u) = 1 - T(u,u,u,u) = 1 - \sum_{i=1}^{u} (u^\top a_j)^4,\ \|u\|_2^2 = 1.$$

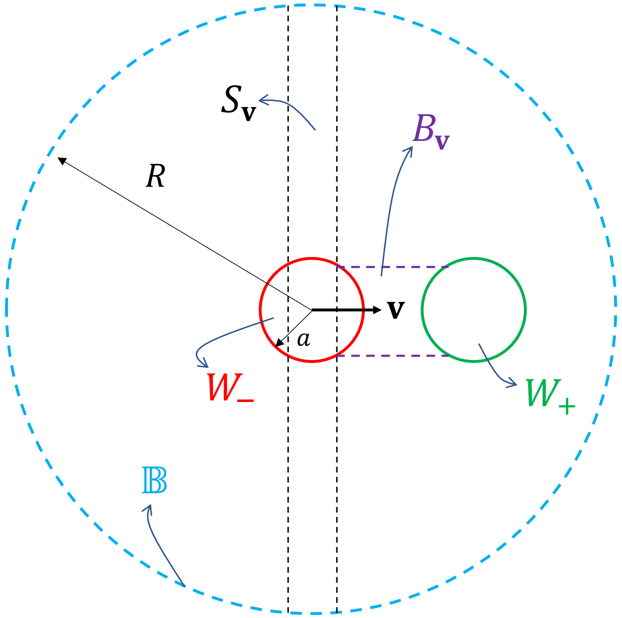
Proposition 1 (Tensor decomposition, informal)

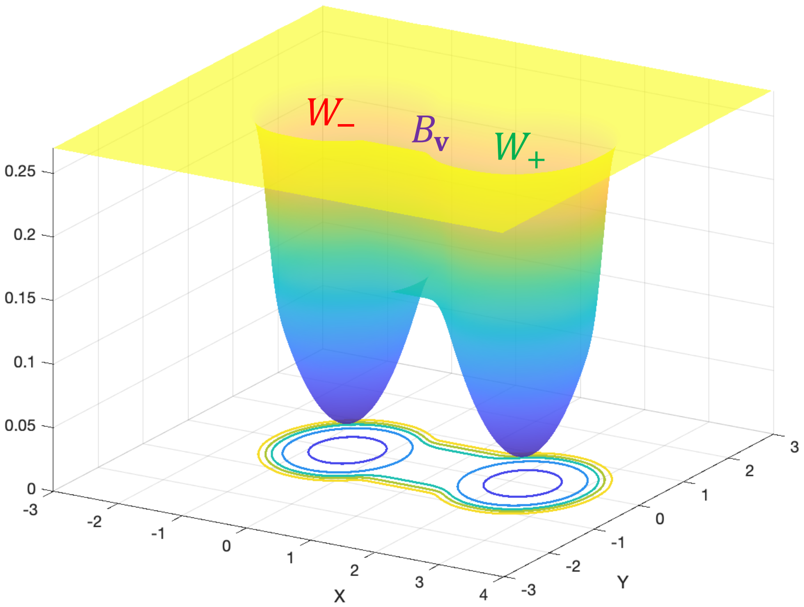
Let d be the dimension of the components of the $\boxed{\text{fourth-order tensor } T \in \mathbb{R}^{d^4}}$, δ be the expected risk yielded by the limit distribution μ_{QTW}, and ϵ be the maximum error between μ_{QTW} and the actual obtained distribution (quantified by L^1 norm). For sufficiently small ϵ and sufficiently small δ, the total time T_{tot} for finding all orthogonal components of T by QTW satisfies

$$T_{\mathrm{tot}} = O(\mathrm{poly}(1/\delta, e^d, 1/\epsilon))e^{\frac{(d-1)+o_\delta(1)}{2\delta}}.$$

Quantum Speedup

Significant speedup by quantum tunneling walks compared to classical algorithms with gradient queries is possible:





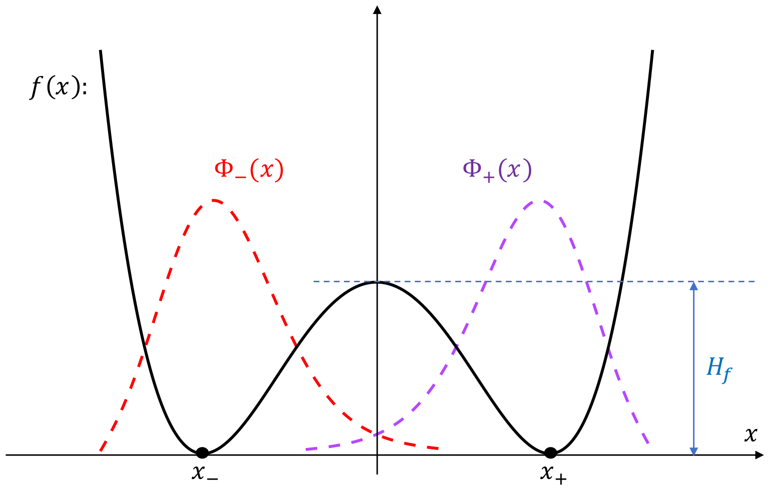
Theorem 2 (Advantage with conditions, informal)

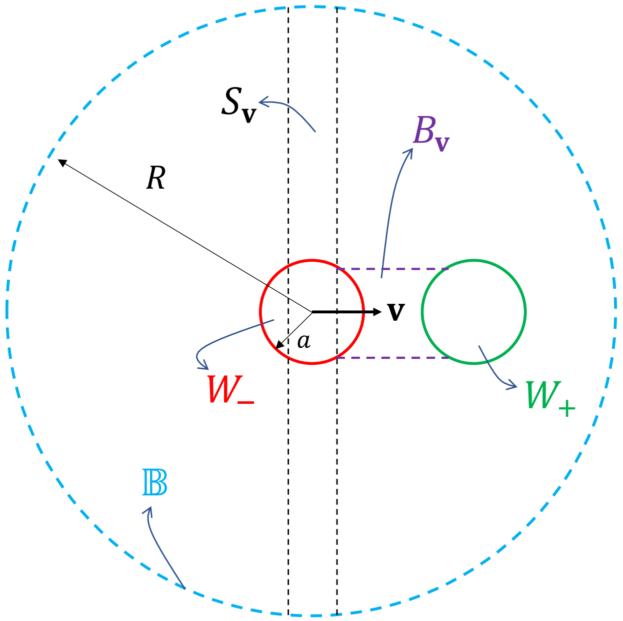
For any dimension d, there exists a landscape f : $\mathbb{R}^d \to \mathbb{R}$ *such that its local minima are global minima, and on which, with high probability, QTW can hit the neighborhood of an unknown global minimum from the local ground state associated to a known minimum using queries polynomial in d, while no classical algorithm knowing the same minimum can hit the same target region with queries subexponential in d.*

Caveat!

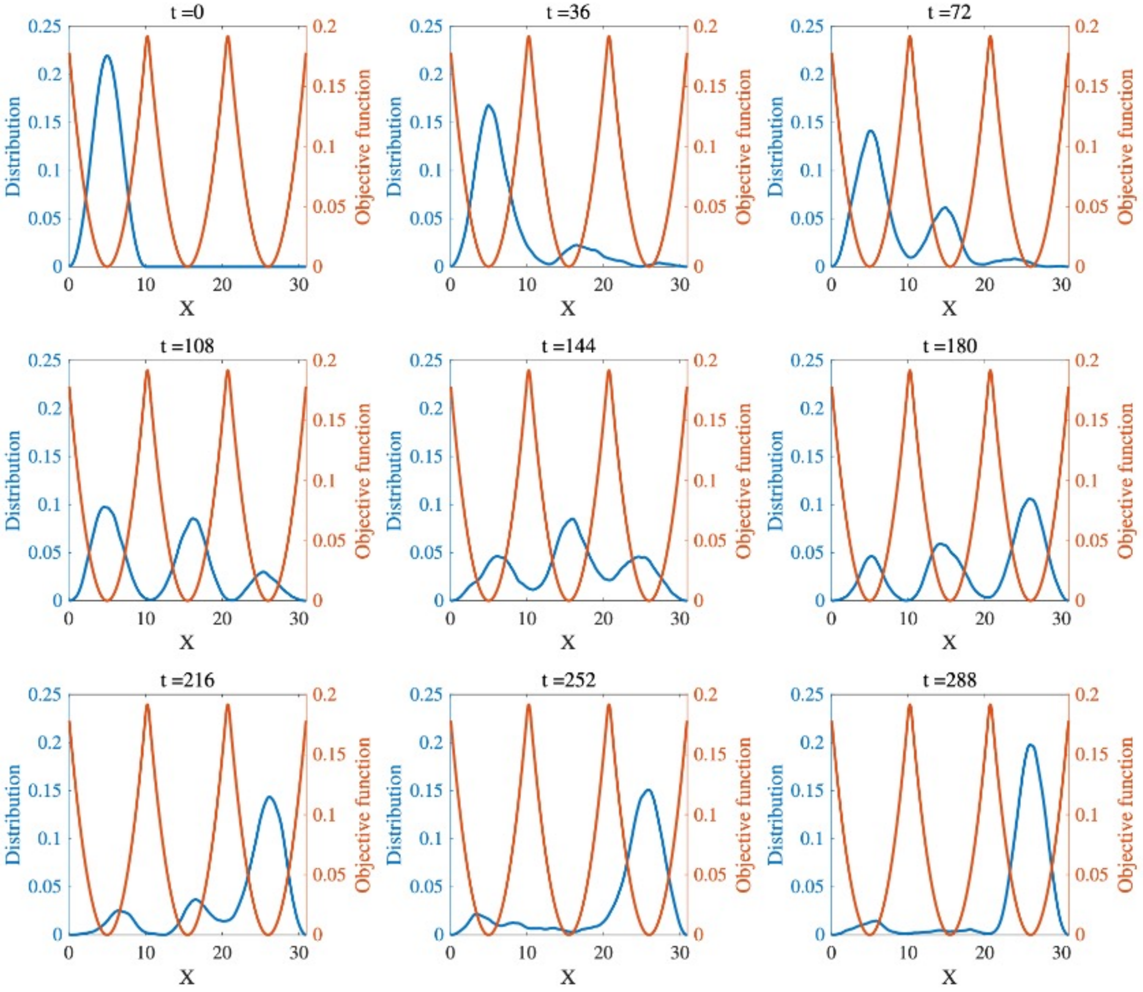
Paradox: This is like an unstructured search problem - reduction from Grover's algorithm?

initial state reveals directional information

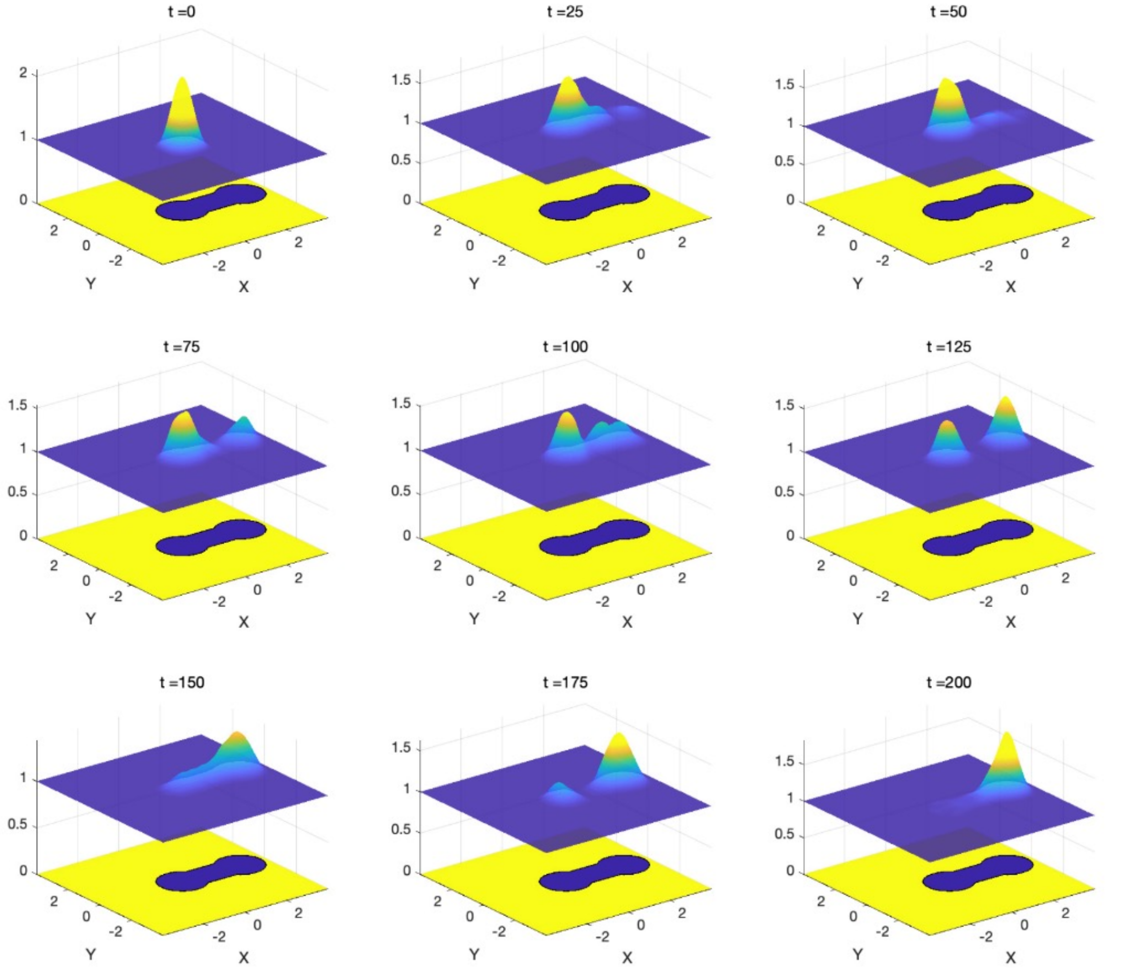




Numerical experiments

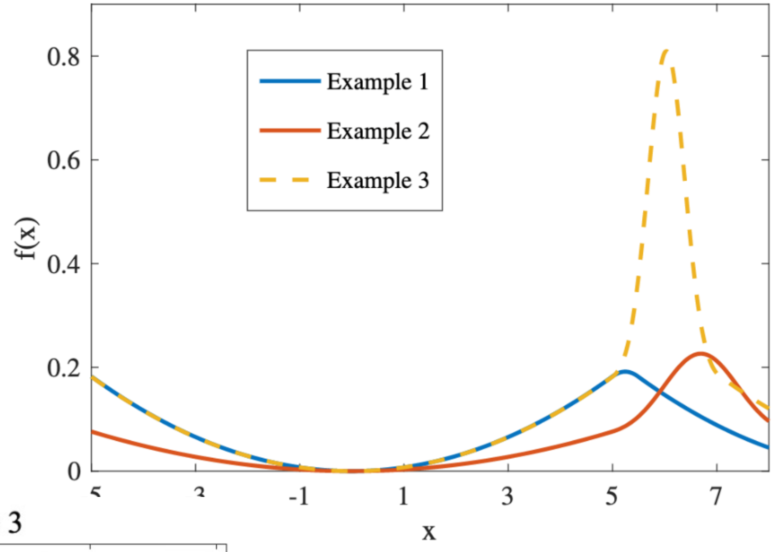


Experiments on three consecutive wells



Experiments on the speedup example (d=2)

Numerical experiments



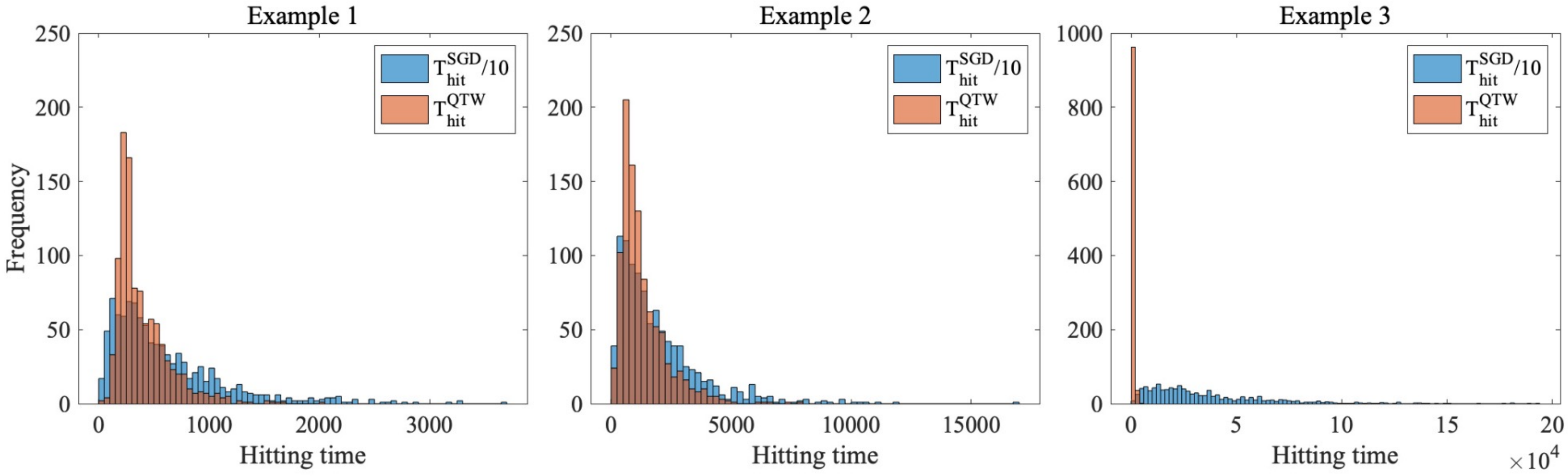


Figure 8: Quantum-classical comparison between SGD and QTW on three landscapes. Example 1 is the critical case where the exponential terms in QTW and SGD evolution time are equal for sufficiently small accuracy δ. Example 2 has flatter minima but similar barriers compared to Example 1, enabling QTW to be faster. Example 3 possesses the same flatness of minima as Example 1 but is equipped with sharp but thin barriers, enabling larger quantum speedups. We take $\tau = 288, 800, 600$ in the three examples, respectively.

Conclusions

A first attempt and framework for quantum algorithms for nonconvex optimization

- The algorithm is simple: quantum simulation of the Schrödinger equation
- Between two local minima: quantum tunneling
- Dynamics among all minima: quantum walks

Open questions:

- ➢ Looser assumptions or more general landscapes?
- ➢ Quantum speedups on more examples, which can give real quantum-classical separation?
- ➢ Other dynamics?

Thank you! Questions: tongyangli@pku.edu.cn