# Quantum Algorithms for Escaping from Saddle Points

Chenyi Zhang[1], Jiaqi Leng[2,3], and Tongyang Li[3,4]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Department of Mathematics, University of Maryland
[3]QuICS and UMIACS, University of Maryland
[4]Center for Theoretical Physics, Massachusetts Institute of Technology

We initiate the study of quantum algorithms for escaping from saddle points with provable guarantee. Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$, our quantum algorithm outputs an $\epsilon$-approximate local minimum using $\tilde{O}(\log^2 n/\epsilon^{1.75})$[1] queries to the quantum evaluation oracle (i.e., the zeroth-order oracle). Compared to the classical state-of-the-art algorithm by Jin et al. with $\tilde{O}(\log^6 n/\epsilon^{1.75})$ queries to the gradient oracle (i.e., the first-order oracle), our quantum algorithm is polynomially better in terms of $n$ and matches its complexity in terms of $1/\epsilon$. Our quantum algorithm is built upon two techniques: First, we replace the classical perturbations in gradient descent methods by simulating quantum wave equations, which constitutes the polynomial speedup in $n$ for escaping from saddle points. Second, we show how to use a quantum gradient computation algorithm due to Jordan to replace the classical gradient queries in nonconvex optimization by quantum evaluation queries with the same complexity, extending the same result from convex optimization due to van Apeldoorn et al. and Chakrabarti et al. Finally, we also perform numerical experiments that support our quantum speedup.

**Motivations.** Nonconvex optimization has been a central research topic in optimization theory in the past decade, mainly because the loss functions in many machine learning models (including neural networks) are typically nonconvex. However, finding the global optima of a nonconvex function is NP-hard in general. Instead, many theoretical works focus on finding local optima, since there are landscape results suggesting that local optima are nearly as good as the global optima for many learning problems (see e.g. [5, 19–22, 25]). On the other hand, it is known that saddle points (and local maxima) can give highly suboptimal solutions in many problems; see e.g. [26, 35]. Furthermore, saddle points are ubiquitous in high-dimensional nonconvex optimization problems [8, 15, 17].

Therefore, one of the most important problems in nonconvex optimization theory is to *escape from saddle points*. Suppose we have a twice-differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that

- $f$ is $\ell$-smooth: $\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \le \ell\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$,
- $f$ is $\rho$-Hessian Lipschitz: $\|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \le \rho\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$;

the goal is to find an $\epsilon$-approximate local minimum $\mathbf{x}_\epsilon$ such that[2]

$$\|\nabla f(\mathbf{x}_\epsilon)\| \le \epsilon, \quad \lambda_{\min}(\nabla^2 f(\mathbf{x}_\epsilon)) \ge -\sqrt{\rho\epsilon}. \tag{1}$$

There have been two main considerations on designing algorithms for escaping from saddle points. First, algorithms with good performance in practice are typically dimension-free or almost dimension-free (i.e., having $\mathrm{poly}(\log n)$ dependence), especially considering that most machine learning models in the real world have enormous dimensions. Second, practical algorithms prefer simple oracle access to the nonconvex function. If we are given a Hessian oracle of $f$, which takes $\mathbf{x}$ as the input and outputs $\nabla^2 f(\mathbf{x})$, we can find an $\epsilon$-approximate local minimum by second-order methods; for instance, Ref. [33] takes $O(1/\epsilon^{1.5})$ queries.

[1]The $\tilde{O}$ notation omits poly-logarithmic terms, i.e., $\tilde{O}(g) = O(g\,\mathrm{poly}(\log g))$.

[2]In general, we can ask for an $(\epsilon_1, \epsilon_2)$-approximate local minimum $\mathbf{x}$ such that $\|\nabla f(\mathbf{x})\| \le \epsilon_1$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\epsilon_2$. The scaling in (1) was first adopted by [33] and has been taken as a standard by subsequent works (e.g. [1, 9, 16, 27–29, 38–40]).

However, because the Hessian is an $n \times n$ matrix, its construction takes $\Omega(n^2)$ cost in general. Therefore, it has become a notable interest to escape from saddle points using simpler oracles.

A seminal work along this line was by Ge et al. [19], which can find an $\epsilon$-approximate local minimum satisfying (1) only using the first-order oracle, i.e., gradients. Although this paper has a $\mathrm{poly}(n)$ dependence in the query complexity of the oracle, the follow-up work [27] can make it almost dimension-free with complexity $\tilde{O}(\log^4 n/\epsilon^2)$, and the state-of-the-art result takes $\tilde{O}(\log^6 n/\epsilon^{1.75})$ queries [29]. However, these results suffer from a significant overhead in terms of $\log n$, and it has been an open question to keep both the merits of using only the first-order oracle as well as being close to dimension-free [30].

In this paper, we explore quantum algorithms for escaping from saddle points. This is a mutual generalization of both classical and quantum algorithms for optimization:

- For quantum computing, the vast majority of previous quantum optimization algorithms had been devoted to convex optimization with the focuses on semidefinite programs [2, 3, 6, 7, 32] and general convex optimization [4, 10]; these results have at least a $\sqrt{n}$ dependence in their complexities, and their quantum algorithms are far from dimension-free methods. Up to now, little is known about quantum algorithms for nonconvex optimization.

  However, there are inspirations that quantum speedups in nonconvex scenarios can potentially be more significant than convex scenarios. In particular, *quantum tunneling* is a phenomenon in quantum mechanics where the wave function of a quantum particle can tunnel through a potential barrier and appear on the other side with significant probability. This very much resembles escaping from poor landscapes in nonconvex optimization. Moreover, quantum algorithms motivated by quantum tunneling will be essentially different from those motivated by the Grover search [24], and will demonstrate significant novelty if the quantum speedup compared to the classical counterparts is more than quadratic.

- For classical optimization theory, since many classical optimization methods are physics-motivated, including Nesterov's momentum-based methods [34], stochastic gradient Langevin dynamics [41] or Hamiltonian Monte Carlo [18], etc., the elevation from classical mechanics to quantum mechanics can potentially bring more observations on designing fast *quantum-inspired classical algorithms*. In fact, quantum-inspired classical machine learning algorithms have been an emerging topic in theoretical computer science [11–13, 23, 36, 37], and it is worthwhile to explore relevant classical algorithms for optimization.

**Contributions.** Our main contribution is a quantum algorithm that finds an $\epsilon$-approximate local minimum with polynomial quantum speedup in $n$ compared to the classical state-of-the-art result [29] using the gradient oracle (the first-order oracle). Furthermore, our quantum algorithm only takes queries to the *quantum evaluation oracle* (the zeroth-order oracle), which is defined as a unitary map $U_f$ on $\mathbb{R}^n \otimes \mathbb{R}$ such that

$$U_f|\mathbf{x}\rangle|\mathbf{0}\rangle = |\mathbf{x}\rangle|f(\mathbf{x})\rangle \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{2}$$

Note that if the classical evaluation oracle can be implemented by explicit arithmetic circuits, the quantum evaluation oracle in (2) can be implemented by quantum arithmetic circuits of the same size up to polylogarithmic factors. As a result, it has been the standard assumption in previous literature on quantum algorithms for convex optimization [4, 10], and we subsequently adopt it here for nonconvex optimization.

**Theorem 1** (Main result, informal)**.** *There is a quantum algorithm that finds an $\epsilon$-approximate local minimum satisfying (1), using $\tilde{O}(\log^2 n/\epsilon^{1.75})$ queries to the quantum evaluation oracle (2).*

Technically, our work is inspired by both the perturbed gradient descent (PGD) algorithm in [27, 28] and the perturbed accelerated gradient descent (PAGD) algorithm in [29]. To be more specific, PGD applies gradient descent iteratively until it reaches a point with small gradient. It can potentially be a saddle point, so PGD applies uniform perturbation in a small ball centered at that point and then continues the GD iterations. It can be shown that with an appropriate choice of the radius, PGD can shake the point off from the saddle

| Reference | Queries | Oracle |
|---|---|---|
| [14, 33] | $O(1/\epsilon^{1.5})$ | Hessian |
| [1, 9] | $\tilde{O}(\log n/\epsilon^{1.75})$ | Hessian-vector product |
| [27, 28] | $\tilde{O}(\log^4 n/\epsilon^2)$ | Gradient |
| [29] | $\tilde{O}(\log^6 n/\epsilon^{1.75})$ | Gradient |
| **this work** | $\tilde{O}(\log^2 n/\epsilon^{1.75})$ | Quantum evaluation |

Table 1: A summary of the state-of-the-art works on finding an approximate local minimum using different oracles. The query complexities are highlighted in terms of the dimension $n$ and the error $\epsilon$.

and converge to a local minimum with high probability. The PAGD in [29] adopts the similar perturbation idea, but the GD is replaced by Nesterov's AGD [34].

Our quantum algorithm is built upon PGD and PAGD and shares their simplicity of being single-loop, but we propose two main modifications. On the one hand, for the perturbation steps for escaping from saddle points, we replace the uniform perturbation by evolving a quantum wave function governed by the Schrödinger equation and using the measurement outcome as the perturbed result. Intuitively, the Schrödinger equation is able to screen the local geometry of a saddle point through wave interference, which results in a phenomenon that the wave packet disperses rapidly along the directions with significant function value decrease. Specifically, quantum mechanics finds the negative curvature directions more efficiently than the classical counterpart: for a constant $\epsilon$, the classical PGD takes $O(\log n)$ steps to decrease the function value by $\Omega(1/\log^3 n)$ with high probability, and the PAGD takes $O(\log n)$ steps to decrease the function value by $\Omega(1/\log^5 n)$ with high probability. Quantumly, the simulation of the Schrödinger equation for time $t$ takes $\tilde{O}(t \log n)$ evaluation queries, but simulation for time $t = O(\log n)$ and $O(\log n)$ subsequent GD iterations suffice to decrease the function value by $\Omega(1)$ with high probability. This is summarized as Table 2 below. In addition, our quantum algorithm is *classical-quantum hybrid*: the transition between consecutive iterations is still classical, while the only quantum computing part happens in each iteration for replacing the classical uniform perturbation.

| | Perturbation | # iterations/ simulation time | Function decrease | Queries in each iteration/unit time |
|---|---|---|---|---|
| Classical | Uniform in ball | $O(\log n)$ | $\Omega(1/\log^5 n)$ | 1 |
| Quantum | Quantum simulation | $O(\log n)$ | $\Omega(1)$ | $\tilde{O}(\log n)$ |

Table 2: A detailed comparison between our result and the classical state-of-the-art result [29], assuming $\epsilon = \Theta(1)$.

On the other hand, for the gradient descent steps, we replace them by a quantum algorithm for computing gradients using also quantum evaluation queries. The idea was initiated by Jordan [31] who computed the gradient at a point by applying the quantum Fourier transform on a mesh near the point. Prior arts have shown how to apply Jordan's algorithm for convex optimization [4, 10], and we conduct a detailed analysis showing how in nonconvex optimization we can replace classical gradient queries by the same number of quantum evaluation queries. Technically, we essentially show the robustness of escaping from saddle points by PGD, which may be of independent interest.

Finally, we perform numerical experiments that support our quantum speedup. Specifically, we observe the dispersion of quantum wave packets along the negative curvature direction in various landscapes. In a comparative study, our PGD with quantum simulation outperforms the classical PGD with a higher probability of escaping from saddle points and fewer iteration steps. We also compare the dimension dependence of classical and quantum algorithms in a model question with dimensions varying from 10 to 1000, and our quantum algorithm achieves a better dimension scaling overall.

## References


[1] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma, *Finding approximate local minima faster than gradient descent*, Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1195–1199, 2017, arXiv:1611.01146.

[2] Joran van Apeldoorn and András Gilyén, *Improvements in quantum SDP-solving with applications*, Proceedings of the 46th International Colloquium on Automata, Languages, and Programming, Leibniz International Proceedings in Informatics (LIPIcs), vol. 132, pp. 99:1–99:15, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, arXiv:1804.05058.

[3] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf, *Quantum SDP-solvers: Better upper and lower bounds*, 58th Annual Symposium on Foundations of Computer Science, IEEE, 2017, arXiv:1705.01843.

[4] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf, *Convex optimization using quantum oracles*, Quantum **4** (2020), 220, arXiv:1809.00643.

[5] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro, *Global optimality of local search for low rank matrix recovery*, Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 3880–3888, 2016, arXiv:1605.07221.

[6] Fernando G.S.L. Brandão, Amir Kalev, Tongyang Li, Cedric Yen-Yu Lin, Krysta M. Svore, and Xiaodi Wu, *Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning*, Proceedings of the 46th International Colloquium on Automata, Languages, and Programming, Leibniz International Proceedings in Informatics (LIPIcs), vol. 132, pp. 27:1–27:14, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, arXiv:1710.02581.

[7] Fernando G.S.L. Brandão and Krysta Svore, *Quantum speed-ups for semidefinite programming*, Proceedings of the 58th Annual Symposium on Foundations of Computer Science, pp. 415–426, 2017, arXiv:1609.05537.

[8] Alan J Bray and David S. Dean, *Statistics of critical points of Gaussian fields on large-dimensional spaces*, Physical review letters **98** (2007), no. 15, 150201, arXiv:cond-mat/0611023.

[9] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford, *Accelerated methods for nonconvex optimization*, SIAM Journal on Optimization **28** (2018), no. 2, 1751–1772, arXiv:1611.00756.

[10] Shouvanik Chakrabarti, Andrew M. Childs, Tongyang Li, and Xiaodi Wu, *Quantum algorithms and lower bounds for convex optimization*, Quantum **4** (2020), 221, arXiv:1809.01731.

[11] Nai-Hui Chia, András Gilyén, Tongyang Li, Han-Hsuan Lin, Ewin Tang, and Chunhao Wang, *Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing quantum machine learning*, Proceedings of the 52nd Annual ACM Symposium on Theory of Computing, pp. 387–400, ACM, 2020, arXiv:1910.06151.

[12] Nai-Hui Chia, Tongyang Li, Han-Hsuan Lin, and Chunhao Wang, *Quantum-inspired sublinear algorithm for solving low-rank semidefinite programming*, To appear in the proceedings of the 45th International Symposium on Mathematical Foundations of Computer Science, 2020, arXiv:1901.03254.

[13] Nai-Hui Chia, Han-Hsuan Lin, and Chunhao Wang, *Quantum-inspired sublinear classical algorithms for solving low-rank linear systems*, 2018, arXiv:1811.04852.

[14] Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi, *A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization*, Mathematical Programming **162** (2017), no. 1-2, 1–32.

[15] Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, Advances in neural information processing systems, pp. 2933–2941, 2014, arXiv:1406.2572.

[16] Cong Fang, Zhouchen Lin, and Tong Zhang, *Sharp analysis for nonconvex SGD escaping from saddle points*, Conference on Learning Theory, pp. 1192–1234, 2019, arXiv:1902.00247.

[17] Yan V. Fyodorov and Ian Williams, *Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity*, Journal of Statistical Physics **129** (2007), no. 5-6, 1081–1116, arXiv:cond-mat/0702601.

[18] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu, *Global convergence of stochastic gradient Hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration*, 2018, arXiv:1809.04618.

[19] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points – online stochastic gradient for tensor decomposition*, Proceedings of the 28th Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 40, pp. 797–842, 2015, arXiv:1503.02101.

[20] Rong Ge, Jason D. Lee, and Tengyu Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems, pp. 2981–2989, 2016, arXiv:1605.07272.

[21] Rong Ge, Jason D. Lee, and Tengyu Ma, *Learning one-hidden-layer neural networks with landscape design*, International Conference on Learning Representations, 2018, arXiv:1711.00501.

[22] Rong Ge and Tengyu Ma, *On the optimization landscape of tensor decompositions*, Advances in Neural Information Processing Systems, pp. 3656–3666, Curran Associates Inc., 2017, arXiv:1706.05598.

[23] András Gilyén, Seth Lloyd, and Ewin Tang, *Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension*, 2018, arXiv:1811.04909.

[24] Lov K. Grover, *A fast quantum mechanical algorithm for database search*, Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing, pp. 212–219, ACM, 1996, arXiv:quant-ph/9605043.

[25] Moritz Hardt, Tengyu Ma, and Benjamin Recht, *Gradient descent learns linear dynamical systems*, Journal of Machine Learning Research **19** (2018), no. 29, 1–44, arXiv:1609.05191.

[26] Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli, *Global convergence of non-convex gradient descent for computing matrix squareroot*, Artificial Intelligence and Statistics, pp. 479–488, 2017, arXiv:1507.05854.

[27] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan, *How to escape saddle points efficiently*, Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1724–1732, 2017, arXiv:1703.00887.

[28] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan, *Stochastic gradient descent escapes saddle points efficiently*, 2019, arXiv:1902.04811.

[29] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, Conference on Learning Theory, pp. 1042–1085, 2018, arXiv:1711.10456.

[30] Michael I. Jordan, *On gradient-based optimization: Accelerated, distributed, asynchronous and stochastic optimization*, 2017, https://www.youtube.com/watch?v=VE2ITg%5FhGnI.

[31] Stephen P. Jordan, *Fast quantum algorithm for numerical gradient estimation*, Physical Review Letters **95** (2005), no. 5, 050501, arXiv:quant-ph/0405146.

[32] Iordanis Kerenidis and Anupam Prakash, *A quantum interior point method for LPs and SDPs*, 2018, arXiv:1808.09266.

[33] Yurii Nesterov and Boris T. Polyak, *Cubic regularization of Newton method and its global performance*, Mathematical Programming **108** (2006), no. 1, 177–205.

[34] Yurii E. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$, Soviet Mathematics Doklady, vol. 27, pp. 372–376, 1983.

[35] Ju Sun, Qing Qu, and John Wright, *A geometric analysis of phase retrieval*, Foundations of Computational Mathematics **18** (2018), no. 5, 1131–1198, arXiv:1602.06664.

[36] Ewin Tang, *Quantum-inspired classical algorithms for principal component analysis and supervised clustering*, 2018, arXiv:1811.00414.

[37] Ewin Tang, *A quantum-inspired classical algorithm for recommendation systems*, Proceedings of the 51st Annual ACM Symposium on Theory of Computing, pp. 217–228, ACM, 2019, arXiv:1807.04271.

[38] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I. Jordan, *Stochastic cubic regularization for fast nonconvex optimization*, Advances in neural Information Processing Systems, pp. 2899–2908, 2018, arXiv:1711.02838.

[39] Yi Xu, Rong Jin, and Tianbao Yang, *NEON+: Accelerated gradient methods for extracting negative curvature for non-convex optimization*, 2017, arXiv:1712.01033.

[40] Yi Xu, Rong Jin, and Tianbao Yang, *First-order stochastic algorithms for escaping from saddle points in almost linear time*, Advances in Neural Information Processing Systems, pp. 5530–5540, 2018, arXiv:1711.01944.

[41] Yuchen Zhang, Percy Liang, and Moses Charikar, *A hitting time analysis of stochastic gradient Langevin dynamics*, Conference on Learning Theory, pp. 1980–2022, 2017, arXiv:1702.05575.