HOUDINI: Escaping from Moderately Constrained Saddles

Dmitrii Avdiukhin DAVDYUKH@IU.EDU
Indiana University, IN

Grigory Yaroslavtsev GRIGORY@GRIGORY.US
George Mason University, VA

Abstract

We give polynomial time algorithms for escaping from high-dimensional saddle points under a moderate number of constraints. Given gradient access to a smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ we show that (noisy) gradient descent methods can escape from saddle points under a logarithmic number of inequality constraints. This constitutes progress (without reliance on NP-oracles or altering the definitions to only account for certain constraints) on the main open question of the breakthrough work of Ge et al. [16] who showed an analogous result for unconstrained and equality-constrained problems. Our results hold for both regular and stochastic gradient descent.

1. Introduction

Achieving convergence of gradient descent to an (approximate) local minimum is a central question in non-convex optimization for machine learning. In recent years, breakthrough progress starting with the work of Ge et al. [16] has led to a flurry of results in this area (see e.g. Carmon and Duchi [9, 10], Carmon et al. [11], Du et al. [14], Jin et al. [19, 20, 21], Mokhtari et al. [23], Staib et al. [31]), culminating in almost optimal bounds [34]. However, despite this success a key open question of [16] still remains unanswered – can gradient methods efficiently escape from saddle points in *constrained* non-convex optimization? In fact, even basic linear inequality constraints still remain an obstacle: "Dealing with inequality constraints is left as future work" Ge et al. [16][1]. This is due to the NP-hardness of the related copositivity problem [24], which corresponds to the case when the number of constraints is linear in the dimension. In this paper, we make progress on this open question in the case when the number of constraints depends moderately on the dimension.

Consider a feasible set defined by k linear inequality constraints: $S = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} \le \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{k\times d}$ and $\mathbf{b} \in \mathbb{R}^k$. Let $\mathcal{B}_d(\mathbf{x}, r)$ be a d-dimensional closed ball of radius r centred at $\mathbf{x}$. We write $\mathcal{B}(\mathbf{x}, r)$ when the dimension is clear from the context and drop the first parameter when $\mathbf{x} = \mathbf{0}$. Our goal is to minimize the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ over S, i.e. $\min_{x\in S} f(x)$. We first introduce standard assumptions of boundedness and smoothness. A function f is *bounded* if it takes values in $[-1, 1]$ which can always be achieved by rescaling.

Assumption 1 (Smoothness) *The objective function f satisfies the following properties:*

1. *(First order) f has an L-Lipschitz gradient (f is L-smooth):* $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
2. *(Second order) f has a ρ-Lipschitz Hessian:* $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \le \rho\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

1. Using Lagrangian multipliers, equality constraints can be seen as reducing the dimension of the otherwise unconstrained problem.

Definition 1 (Local minimum) *For $S \subseteq \mathbb{R}^d$, let $f : \mathbb{R}^d \to \mathbb{R}$. A point $\mathbf{x}^\star$ is a local minimum of f in S if and only if there exists $r > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x} \in S \cap \mathcal{B}(\mathbf{x}^\star, r)$.*

Since finding a local minimum is NP-hard even in the unconstrained case (see e.g. Anandkumar and Ge [3] and the references within), the notion of a local minimum is typically relaxed, a natural relaxation being the following:

Definition 2 (Approximate local minimum) *For $S \subseteq \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ a point $\mathbf{x}^\star$ is a (δ, r)-approximate local minimum if $f(\mathbf{x}) \geq f(\mathbf{x}^\star) - \delta$ for all $\mathbf{x} \in S \cap \mathcal{B}(\mathbf{x}^\star, r)$.*

For smooth functions, a common approach is to instead define stationary points in terms of the gradient and the eigenvalues, which, as shown below, is motivated by the Taylor expansion of f:

Definition 3 ([21, 25]) *A point $\mathbf{x}$ is an ε-second-order stationary point (ε-SOSP) if $\|\nabla f(\mathbf{x})\| < \varepsilon$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x})) > -\sqrt{\rho\varepsilon}$, where $\lambda_{\min}$ denotes the smallest eigenvalue of the Hessian.*

When applying this definition to the constrained case, eigenvectors and eigenvalues are not well-defined, since, even when an escaping direction exists, there might be no eigenvectors inside the feasible set. Moreover, for $f(\mathbf{x}) = -\frac{1}{2}\|\mathbf{x}\|^2$ and any compact feasible set, the Hessian is $-I$ at any point with $\lambda_{\min}(-I) = -1$. Hence an ε-SOSP doesn't exist according to the Definition 3, even though a local minimum exists. In fact, Definition 3 arises from the Taylor expansion, which justifies the choice of $\sqrt{\rho\epsilon}$ as the bound on the smallest eigenvalue. If the function has a ρ-Lipschitz Hessian:

$$\left| f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \mathbf{h}^\top \nabla f(\mathbf{x}) - \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h} \right| \leq \frac{\rho}{6}\|\mathbf{h}\|^3.$$

To guarantee that the discrepancy between the function and its quadratic approximation is small relative to δ (from Definition 2), a natural choice of r is $\sqrt[3]{\delta/\rho}$, which bounds the discrepancy with $\Theta(\delta)$. Therefore, based on the quadratic approximation, one can distinguish a (δ, r)-approximate local minimum from not a $(c\delta, r)$-approximate local minimum for $c < 1$. By setting this r and selecting $\varepsilon = \sqrt[3]{\delta^2\rho}$, we have $\sqrt{\rho\varepsilon} = \sqrt[3]{\delta\rho^2}$ and for any $\mathbf{h} \in \mathcal{B}(\mathbf{x}, r)$ (see Appendix A):

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) \geq \mathbf{h}^\top \nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x})\mathbf{h} - \frac{\rho}{6}\|\mathbf{h}\|^3 \geq -2\delta.$$

Using the ball radius discussed above we arrive at the following version of Definition 2:

Definition 4 (Approximate SOSP) *For $S \subseteq \mathbb{R}^d$, let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function with a ρ-Lipschitz Hessian. A point $\mathbf{x}^\star$ is a δ-second-order stationary point (δ-SOSP) if for $r = \sqrt[3]{\delta/\rho}$:*

$$\inf_{\mathbf{x} \in S \cap \mathcal{B}(\mathbf{x}^\star, r)} f(\mathbf{x}) \geq f(\mathbf{x}^\star) - \delta$$

1.1. Our Results

Our results hold for stochastic gradient descent (SGD) under the following standard assumption:

Assumption 2 *Access to a* stochastic gradient oracle $\mathbf{g}(\mathbf{x})$*:*

1. *(Unbiased expectation)* $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x})$.
2. *(Variance)* $\mathbb{E}[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2] \leq \sigma^2$.

Our main result is the following theorem which quantifies the complexity of finding an approximate SOSP under a moderate number of linear inequality constraints, showing that this problem is solvable in polynomial time for $k = O(\log d)$. We refer to a function as (L, ρ)-smooth if it satisfies Assumption 1 and simply *second-order smooth* if both smoothness parameters are constant.

Theorem 5 (Theorem 8) *Let S be a set defined by an intersection of k linear inequality constraints. Let f be a second-order smooth bounded function. Given access to a stochastic gradient oracle satisfying Assumption 2, there exists an algorithm which for any $\delta > 0$ finds a δ-SOSP in $\tilde{O}(\frac{1}{\delta}(d^3 2^k + \frac{d^2\sigma^2}{\delta^{4/3}}))$ time using $\tilde{O}(\frac{1}{\delta}(d + \frac{d^3\sigma^2}{\delta^{4/3}}))$ stochastic gradient oracle calls. In the deterministic gradient case ($\sigma = 0$), the time complexity is $\tilde{O}(\frac{d^3 2^k}{\delta})$ and the number of gradient oracle calls is $\tilde{O}(\frac{d}{\delta})$.*

The exponential dependence of time complexity on k in our results (not required in the oracle calls) is most likely unavoidable due to the following hardness result, which implies that when $k = d$ then the complexity of this problem can't be polynomial in d under the standard hardness assumptions.

Remark 6 (Matrix copositivity [24]) *For a quadratic function $f(\mathbf{x}) = \mathbf{x}^T \mathbf{M} \mathbf{x}$ subject to constraints $x_i \geq 0$ for all i, it is NP-hard to decide whether there exists a solution with $f(\mathbf{x}) > 0$.*

1.2. Related Work

Related results in convex optimization are covered in Boyd and Vandenberghe [7], Bubeck et al. [8]. Among related results in non-convex optimization here we only focus on the algorithms which only use gradient information.

Unconstrained optimization Recall that an *ϵ-first-order stationary point* (ϵ-FOSP) is defined so that $\|\nabla f(\mathbf{x})\| \leq \epsilon$. Analyses of convergence to an ε-FOSP are a cornerstone of non-convex optimization (see e.g. classic texts Bertsekas [5], Nocedal and Wright [26]). Quantitative analysis of convergence to an ϵ-SOSP (Definition 3) started with the breakthrough work by Ge et al. [16] further refined in Carmon and Duchi [9, 10], Jin et al. [19, 20, 21] and most recently in Zhang and Li [34], who show an almost optimal bound. Due to the prevalence of SGD in deep learning, stochastic methods have attracted the most attention (see Allen-Zhu [1], Allen-Zhu and Li [2], Fang et al. [15], Tripuraneni et al. [32], Xu et al. [33], Zhou and Gu [35], Zhou et al. [36] for the case of Lipschitz gradients and Daneshmand et al. [13], Ge et al. [16] for non-Lipschitz gradients). For an in-depth summary of the previous work on unconstrained non-convex optimization we refer the reader to Jin et al. [21].

Constrained optimization The case of equality constraints is typically reducible to the unconstrained case by using Lagrangian multipliers (see e.g. Ge et al. [16]). However, the general constrained case is substantially more challenging since even the definitions of stationarity require a substantial revision. For first-order convergence a rich literature exists, covering projected gradient, Frank-Wolfe, cubic regularization, etc (see e.g. Mokhtari et al. [23] and the references within). For second-order convergence, the landscape of existing work is substantially sparser due to NP-hardness (Remark 6, [24]). A large body of work focuses on achieving convergence using various forms of NP-oracles (see e.g. Bian et al. [6], Cartis et al. [12], Haeser et al. [17], Mokhtari et al. [23], Nouiehed and Razaviyayn [27]), while another approach is to define stationarity in terms of tight constraints only [4, 22].

Relationship with other definitions of SOSP As discussed in Remark 6, second-order constrained optimization is NP-hard due to the hardness of the matrix copositivity problem. Definitions of constrained SOSP in the previous work fall into two categories: 1) definitions that only consider active constraints ("active constraints only" definitions), 2) definitions that preserve the NP-hardness of the problem and rely on NP-oracles to achieve polynomial-time convergence:

1. ("Active constraints only") In Avdiukhin et al. [4], Lu et al. [22] definitions analogous to Definition 3, and the second-order conditions are given with respect to the set of *active* (i.e. tight for the current iterate) constraints. This allows bypassing the NP-hardness since the point at which the hardness of the copositivity problem applies now becomes a stationary point by definition.

2. (NP-hard) In the results relying on NP-oracles (e.g. [6, 17, 23, 27]) the complexity is shifted to solving black-box quadratic optimization problems of a certain type. A key advantage of these types of approaches is that they can handle an arbitrary number of constraints and hence promising in certain machine learning applications.

What is currently lacking in the state of the art is a quantitative analysis of the complexity of convergence to a second-order stationary point, which shows full dependence on both the dimension and accuracy while defining stationarity with respect to the full set of constraints, instead of just active constraints only[2]. Our goal in Theorem 5 is to address this gap and give such an analysis.

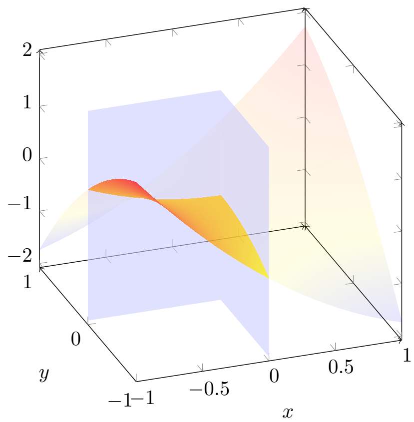
2. Main Result

We address the NP-hardness of the copositivity problem by focusing on the case of a moderate number of constraints and arguing that it can be addressed using gradient-based methods. In order to streamline the presentation, we first focus in Section 2.1 on the key challenge of escaping from a saddle point in a corner defined by the constraints when the underlying function is simply quadratic (see details in Section B). This is already enough to capture some of the key contributions. Our general result is presented in Section 2.2, while more technical details and the full algorithm are given in Appendix C.

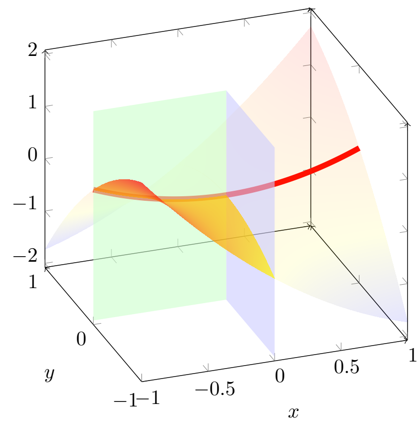
2.1. Quadratic corner saddle point

In this simplified scenario, we consider the case when the function is quadratic without a linear term, and the saddle point is located in the corner of the constrained space. The NP-hardness comes from the fact that the point we aim to find can lie in the intersection of an arbitrary subset of constraints. By doing an exhaustive search over this set of constraints (Algorithm 1) and enforcing them throughout the search process we are able to reduce to a setting similar to the equality-constrained case (Algorithm 2). We show different subsets of constraints enforced by the algorithm

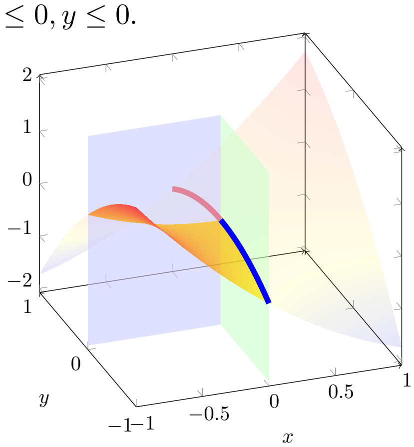
2. While unpublished manuscripts ([18] and [28], relying on [18] as a subroutine) do contain related results, our work differs in a number of important aspects. [18] and [28] require access to the exact gradient and Hessian. We only assume access to the stochastic gradient oracle. Furthermore, [18] assumes access to matrix diagonalization. However, the diagonalization can only be found approximately, which compromises the stability of this approach, especially with respect to the linear term transformations. Our approach handles this issue by appropriate perturbation and using known results for matrix diagonalization. Finally, compared with [18], our main contribution is focused on solving a substantially different problem. While [18] find a global minimum of a quadratic problem, we find an approximate local minimum of an arbitrary smooth non-convex function.



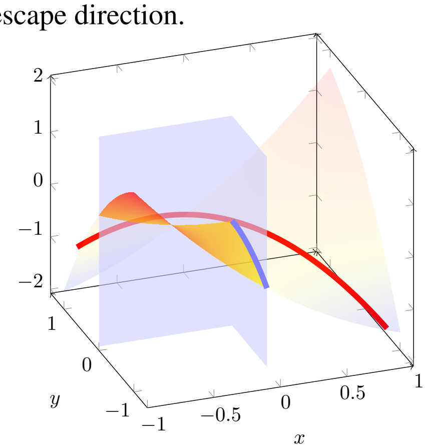
(a) Feasible set with two inequality constraints: $x \leq 0, y \leq 0$.



(b) Active constraint $y = 0$ (green) with no escape direction.



(c) Active constraint $x = 0$ (green) with an escape direction $(0, -1)$. As shown in Lemma 15, out of two escape directions $(0, -1)$ and $(0, 1)$ in this constraint, at least one $(0, -1)$ (blue) lies in the feasible set, and is found by the algorithm.



(d) When no constraints are active, the algorithm finds red directions (negative eigenvectors) outside the feasible set. Note that escape directions with no active constraints exist (blue). Lemma 15 guarantees that we find them in some constrained space (Figure 1(c)).

Figure 1: Escaping quadratic saddle for function $f(x, y) = (\frac{\sqrt{3}}{2}x + \frac{1}{2}y)^2 - (-\frac{1}{2}x + \frac{\sqrt{3}}{2}y)^2$ under constraints $x, y \leq 0$. f is a composition of $g(x, y) = x^2 - y^2$ and rotation by $\pi/6$

in an example in Figure 1. The key challenge is making this argument formal and arguing that this process converges to a constrained approximate SOSP as in Definition 4. This relies on performing a robust analysis of the properties of the points which lead to the hardness of copositivity. In particular, we show that after we guess the set of constraints correctly, the problem reduces to finding the smallest eigenvector (Lemma 14). An exact error analysis of the eigenvector process (Lemma 15, Lemma 16) is then required to complete the proof of the main theorem for this case.

Theorem 7 (Quadratic Corner Saddle Point Escape) *Let $\delta, r > 0$. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ with $\lambda_{|max|}(\mathbf{M}) \le L$ and let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{0}\}$ be defined by k linear inequalies. If $\min_{\mathbf{x}\in\mathcal{B}(r)\cap S} f(\mathbf{x}) < -\delta$, then there exists an Algorithm which with probability at least $1-\xi$ finds a point $\mathbf{x} \in S \cap \mathcal{B}(r)$ with $f(\mathbf{x}) < -\Omega(\delta)$ using $O\left(\frac{Lr^2k2^k}{\delta}\log\frac{1}{\xi}\right)$ deterministic gradient oracle calls.*

The full proof of Theorem 7 is presented in Appendix B.

2.2. General case

In the algorithm for the general case, we address the assumptions made in the quadratic corner saddle point case, while also handling the stochasticity in the gradient. The latter part is standard and is handled via variance reduction in Appendix C.3. The full algorithm iterates the escape subroutine (Algorithm 3) until an escaping point is found. The escape subroutine first approximates the Hessian matrix using the gradient oracle and then performs an exhaustive search over the set of active constraints at the escaping point in a way similar to the quadratic corner case. After the correct subset of constraints is fixed, the current iterate needs to be projected on this set of constraints, which also necessitates a recomputation of various related parameters. When this is done, Algorithm 4 (analogous to Algorithm 2 from the quadratic corner case) finds an escape direction for the resulting quadratic function.

The algorithm tries to find an escaping direction within a ball of radius $r = \sqrt[3]{\delta/\rho}$ (from Definition 4), where the function is well approximated by a quadratic one (by the smoothness assumption, as discussed above). However, compared with the previous section, the function can have a non-negligible linear term – either due to the gradient or as a result of projection on the affine plane – which significantly complicates the algorithm and analysis. We consider the following cases:

Case 1. When the linear term is very large, then either the optimum lies on the boundary of the feasible set S (which is not possible in the case when we guessed the active constraints correctly) or the linear term dominates the quadratic term and is sufficient to escape from a saddle point. In particular, this case covers the situation when the objective is sensitive to the change in the argument (the quadratic term is not sensitive due to the smoothness condition, which bounds its largest eigenvalue).

Case 2. When the optimum lies in the interior of the ball of radius r, none of the constraints are active, and hence we can find the unique unconstrained critical point directly: we find a point where the gradient is zero, which only requires solving a linear system.

Case 3. When the optimum lies on the boundary of the ball of radius r, the only active constraint is $c(\mathbf{y}) = \|\mathbf{y}-\mathbf{x}\|^2 - r^2 = 0$. By KKT conditions, for $\mathbf{y}$ to be the minimizer, there must exist μ such that $\nabla f(\mathbf{y}) = \mu \nabla c(\mathbf{y})$. We show that for each μ, there exists a unique $\mathbf{y}(\mu)$ satisfying the condition above, and only for $O(d)$ values of μ, $\mathbf{y}(\mu)$ lies on the boundary of the ball, resulting in $O(d)$ candidate solutions.

To find these candidates, we first diagonalize the matrix using an orthogonal transformation. This gives a quadratic function whose critical points on the boundary of the ball can be found explicitly (up to the required precision) as roots of certain polynomials. Here we note that the diagonalization performed in this case is the most computationally expensive step in the algorithm, resulting in polynomial dependence on the dimension. We formalize our main result in the following theorem:

Theorem 8 *Let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ be a set defined by an intersection of k linear inequality constraints. Let f satisfy Assumptions 1 and 2 and let $\min_{\mathbf{x}\in S} f(\mathbf{x}) = f^\star$. Then there exists an* *algorithm which for* $\delta > 0$ *finds a* δ*-SOSP in* $\tilde{O}(\frac{f(\mathbf{x}_0)-f^\star}{\delta} d^3 (2^k + \frac{\sigma^2}{\delta^{4/3}}))$ *time using* $\tilde{O}(\frac{f(\mathbf{x}_0)-f^\star}{\delta}(d + \frac{d^3\sigma^2}{\delta^{4/3}}))$ *stochastic gradient oracle calls.*

Next, we outline the proof of Theorem 8. If $\mathbf{x}$ is not a δ-SOSP, our algorithm finds a point $\mathbf{y} \in S \cap \mathcal{B}(\mathbf{x}, r)$ which significantly decreases function value: $f(\mathbf{y}) < f(\mathbf{x}) - \Omega(\delta)$. Therefore, if $\mathbf{x}_0$ is the initial point, our algorithm requires $O(\frac{f(\mathbf{x}_0)-f^\star}{\delta})$ iterations.

As we show in Appendix C.1, we can consider a quadratic approximation of the objective. By guessing which constraints are active at the minimizer $\mathbf{x}^\star$ and enforcing these constraints, we restrict the function to some affine subspace $\mathcal{A}$. By parameterizing $\mathcal{A}$, we eliminate enforced constraints, and, since the rest of the constraints are not active at $\mathbf{x}^\star$, we need to optimize a quadratic function in the intersection of a ball and linear inequality constraints. The full proof is presented in Appendix C.

In conclusion, we prove that, given a logarithmic number of constraints, it's possible to escape from constrained saddle points in polynomial time using stochastic gradient oracles.

References

[1] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2680–2691, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/79a49b3e3762632813f9e35f4ba53d6c-Abstract.html.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. NEON2: finding local minima via first-order oracles. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3720–3730, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d4b2aeb2453bdadaa45cbe9882ffefcf-Abstract.html.

[3] Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 81–102. JMLR.org, 2016. URL http://proceedings.mlr.press/v49/anandkumar16.html.

[4] Dmitrii Avdiukhin, Chi Jin, and Grigory Yaroslavtsev. Escaping saddle points with inequality constraints via noisy sticky projected gradient descent. In *Optimization for Machine Learning Workshop*, 2019.

[5] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48 (3):334–334, 1997.

[6] Wei Bian, Xiaojun Chen, and Yinyu Ye. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1):301–327, 2015.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[9] Yair Carmon and John C. Duchi. Analysis of krylov subspace solutions of regularized non-convex quadratic problems. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10728–10738, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/349f36aa789af083b8e26839bd498af9-Abstract.html.

[10] Yair Carmon and John C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Rev.*, 62(2):395–436, 2020. doi: 10.1137/20M1321759. URL https://doi.org/10.1137/20M1321759.

[11] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR, 2017. URL http://proceedings.mlr.press/v70/carmon17a.html.

[12] Coralia Cartis, Nick IM Gould, and Philippe L Toint. Second-order optimality and beyond: Characterization and evaluation complexity in convexly constrained nonlinear optimization. *Foundations of Computational Mathematics*, 18(5):1073–1107, 2018.

[13] Hadi Daneshmand, Jonas Moritz Kohler, Aurélien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1163–1172. PMLR, 2018. URL http://proceedings.mlr.press/v80/daneshmand18a.html.

[14] Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Aarti Singh, and Barnabás Póczos. Gradient descent can take exponential time to escape saddle points. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1067–1077, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f79921bbae40a577928b76d2fc3edc2a-Abstract.html.

[15] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference*

on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 687–697, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html.

[16] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 797–842. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Ge15.html.

[17] Gabriel Haeser, Hongcheng Liu, and Yinyu Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Mathematical Programming*, 178(1):263–299, 2019.

[18] Yong Hsia and Ruey-Lin Sheu. Trust region subproblem with a fixed number of additional linear inequality constraints has polynomial complexity. *arXiv preprint arXiv:1312.1398*, 2018.

[19] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732. PMLR, 2017. URL http://proceedings.mlr.press/v70/jin17a.html.

[20] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1042–1085. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/jin18a.html.

[21] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *J. ACM*, 68(2): 11:1–11:29, 2021. doi: 10.1145/3418526. URL https://doi.org/10.1145/3418526.

[22] Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1da546f25222c1ee710cf7e2f7a3ff0c-Abstract.html.

[23] Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie. Escaping saddle points in constrained optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/069654d5ce089c13f642d19f09a3d1c0-Paper.pdf.

[24] Katta G. Murty and Santosh N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[25] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[26] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[27] Maher Nouiehed and Meisam Razaviyayn. A trust region method for finding second-order stationarity in linearly constrained nonconvex optimization. *SIAM J. Optim.*, 30(3):2501–2529, 2020. doi: 10.1137/19M1256415. URL https://doi.org/10.1137/19M1256415.

[28] Maher Nouiehed, Jason D Lee, and Meisam Razaviyayn. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2020.

[29] V. Pan. Sequential and parallel complexity of approximate evaluation of polynomial zeros. *Computers & Mathematics with Applications*, 14(8):591–622, 1987. ISSN 0898-1221. doi: https://doi.org/10.1016/0898-1221(87)90186-6. URL https://www.sciencedirect.com/science/article/pii/0898122187901866.

[30] Victor Y. Pan and Zhao Q. Chen. The complexity of the matrix eigenproblem. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99, page 507–516, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130678. doi: 10.1145/301250.301389. URL https://doi.org/10.1145/301250.301389.

[31] Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra. Escaping saddle points with adaptive gradient methods. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5956–5965. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/staib19a.html.

[32] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2904–2913, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/db1915052d15f7815c8b88e879465a1e-Abstract.html.

[33] Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5535–5545, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/217e342fc01668b10cb1188d40d3370e-Abstract.html.

[34] Chenyi Zhang and Tongyang Li. Escape saddle points by a simple gradient-descent based algorithm. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8545–8556. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/47bd8ac1becf213f155a82244b4a696a-Paper.pdf.

[35] Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 3980–3990. PMLR, 2020. URL http://proceedings.mlr.press/v108/zhou20a.html.

[36] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *J. Mach. Learn. Res.*, 21:103:1–103:63, 2020. URL http://jmlr.org/papers/v21/18-447.html.

Appendix A. Notation and Standard Facts

Notation For a set S let $\mathrm{Int}\, S$ be its interior and ∂S be its boundary. For $\mathbf{x} \in \mathbb{R}^d$ and $S \subseteq \mathbb{R}^d$, $\mathrm{Proj}_S(\mathbf{x}) = \mathrm{argmin}_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|$ is the projection of $\mathbf{x}$ on S. For a square matrix $\mathbf{M}$ with eigenvalues $\lambda_1 \le \ldots \le \lambda_d$, we denote $\lambda_{\min}(\mathbf{M}) = \lambda_1$ and $\lambda_{|max|}(\mathbf{M}) = \max(|\lambda_1|, |\lambda_d|)$. For $S = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ and $\mathbf{x} \in S$, we say that i-th constraint is active at $\mathbf{x}$ if $\mathbf{A}_i^\top \mathbf{x} = b_i$, where $\mathbf{A}_i$ is the i-th row of $\mathbf{A}$. $\tilde{O}$ notation hides polylogarithmic dependence on all parameters, including error probability.

Fact 9 (Efficient projection on a polyhedron) *For an arbitrary point $\mathbf{x} \in \mathbb{R}^d$ and a set $S = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ – an intersection of k linear inequality constraints, we can find a projection of $\mathbf{x}$ onto S in time* $\mathrm{poly}(d)$.

Fact 10 (Convergence of power iteration) *Let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a PSD matrix with largest eigenvalue $\lambda_{|max|}$. Let a sequence $\{\mathbf{x}_t\}_{t=0}^\infty$ in $\mathbb{R}^d$ be defined according to the power iteration: $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathbf{x}_{t+1} = \mathbf{M}\mathbf{x}_t$. Then for $\delta < \lambda_{|max|}$ we have $\frac{\mathbf{x}_T^\top \mathbf{M} \mathbf{x}_T}{\|\mathbf{x}_T\|^2} > \delta$ with probability $1 - O(\xi)$ after $T = O\left(\frac{\log(d\delta) - \log(\xi(\lambda_{|max|} - \delta))}{\log \lambda_{|max|}/\delta}\right)$ iterations.*

Proof If $\lambda_i < \delta$, the corresponding component of the vector decreases exponentially with the rate of at least $\delta/\lambda_{|max|}$ and is eventually dominated by the component corresponding to $\lambda_{|max|}$. This leaves us only with components corresponding to $\lambda_i \ge \delta$, finishing the proof.

Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\mathbf{M}$ and let $\mathbf{e}_1, \ldots, \mathbf{e}_d$ be the corresponding eigenvectors. Then $\mathbf{x}_0 = \sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle \mathbf{e}_i$ and $\mathbf{x}_T = \sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle \lambda_i^T \mathbf{e}_i$. Hence,

$$\begin{aligned}
\frac{\mathbf{x}_T^\top \mathbf{M} \mathbf{x}_T}{\|\mathbf{x}_T\|^2} &= \frac{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 \lambda_i^{2T} \lambda_i}{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 \lambda_i^{2T}} \\
&= \frac{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 (\lambda_i/\lambda_{|max|})^{2T} \lambda_i}{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 (\lambda_i/\lambda_{|max|})^{2T}} \\
&= \delta + \frac{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 (\lambda_i/\lambda_{|max|})^{2T} (\lambda_i - \delta)}{\sum_{i=1}^d \langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 (\lambda_i/\lambda_{|max|})^{2T}}.
\end{aligned}$$

It suffices to show that the numerator in the last term is non-negative. Clearly, in the summation, the terms with $\lambda_i \ge \delta$ are positive and the terms with $\lambda_i < \delta$ are negative. Hence, we must lower-bound $\langle \mathbf{x}_0, \mathbf{e}_i \rangle^2$ when $\lambda_i \ge \delta$ and upper-bound $\langle \mathbf{x}_0, \mathbf{e}_i \rangle^2$ when $\lambda_i < \delta$. Since $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, I)$, the following holds for all $i \in [d]$:

- With probability $1 - O(\xi/d)$, $|\langle \mathbf{x}_0, \mathbf{e}_i \rangle| \ge \xi/d$.
- With probability $1 - O(\xi/d)$, $|\langle \mathbf{x}_0, \mathbf{e}_i \rangle| \le \log d/\xi$.

By the union bound, these inequalities hold for all i with probability $1 - O(\xi)$. For i such that $\lambda_i = \lambda_{|max|}$, we have

$$\langle \mathbf{x}_0, \mathbf{e}_d \rangle^2 (\lambda_i/\lambda_{|max|})^{2T} (\lambda_{|max|} - \delta) \ge (\xi/d)^2 (\lambda_{|max|} - \delta).$$

For all i such that $\lambda_i < \delta$, we have:

$$\langle \mathbf{x}_0, \mathbf{e}_i \rangle^2 (\lambda_i/\lambda_{|max|})^{2T} (\lambda_i - \delta) \ge -\delta(\delta/\lambda_{|max|})^{2T} \log^2 d/\xi.$$

Since there are at most d such terms, to guarantee that $\sum_{i=1}^{d}\langle \mathbf{x}_0, \mathbf{e}_i\rangle^2 ({\lambda_i}/{\lambda_{|max|}})^{2T}(\lambda_i - \delta) \geq 0$, it suffices to have:

$$({\xi}/{d})^2(\lambda_{|max|} - \delta) \geq d\delta({\delta}/{\lambda_{|max|}})^{2T} \log^2 {d}/{\xi},$$

which holds when

$$\begin{aligned} T &\geq \frac{\log\left(\frac{d\delta}{(\lambda_{|max|}-\delta)}({d}/{\xi})^2 \log^2 {d}/{\xi}\right)}{2\log {\lambda_{|max|}}/{\delta}} \\ &= O\left(\frac{\log(d\delta) - \log(\xi(\lambda_{|max|} - \delta))}{\log {\lambda_{|max|}}/{\delta}}\right). \end{aligned}$$

■

Fact 11 (Coordinate-wise median trick) *Let $\xi^{(1)}, \dots, \xi^{(n)}$ be independent random variables in $\mathbb{R}^d$ sampled from the same distribution with mean $\mathbb{E}[\xi^{(i)}] = \mu$ and variance $\mathbb{E}[\|\xi^{(1)} - \mu\|^2] \leq \sigma^2$. If ζ is a coordinate-wise median of $\xi_1, \dots, \xi_n$, then*

$$\Pr[\|\zeta - \mu\| > 2\sigma] \leq d \cdot e^{-\Omega(n)}.$$

Proof The proof idea is to apply the median trick to each coordinate separately. For each $\xi^{(i)}$ we have:

$$\mathbb{E}[\|\xi^{(i)} - \mu\|^2] = \mathbb{E}[\sum_{j=1}^{d}(\xi_j^{(i)} - \mu_j)^2] = \sum_{j=1}^{d}\mathbb{E}[(\xi_j^{(i)} - \mu_j)^2].$$

For each coordinate j, let $\sigma_j^2 = \mathbb{E}[(\xi_j^{(i)} - \mu_j)^2]$. Using Chebyshev's inequality, we have:

$$\Pr[(\xi_j^{(i)} - \mu_j)^2 \geq 4\sigma_j^2] \leq \frac{1}{4}.$$

Let $X_{i,j}$ be random variables such that $X_{i,j} = 1$ if $(\xi_j^{(i)} - \mu_j)^2 \geq 4\sigma_j^2$ and $X_{i,j} = 0$ otherwise. From the above inequality, we have $\mathbb{E}[\sum_{i=1}^{n} X_{i,j}] \leq \frac{n}{4}$. Since $X_{1,j}, \dots, X_{n,j}$ are bounded and independent, by the Chernoff bound we have:

$$\Pr[\sum_{i=1}^{n} X_{i,j} > \frac{n}{2}] = e^{-\Omega(n)}.$$

Hence, for any j, with probability $1 - e^{-\Omega(n)}$ at most half of $X_{i,j}$ are 1. Therefore, at least half of $\xi_j^{(1)}, \dots, \xi_j^{(n)}$ lie in $[\mu_j - \sigma_j, \mu_j + \sigma_j]$, and hence the same holds for their median ζ_j. Taking the union bound over all coordinates, with probability $1 - d \cdot e^{-\Omega(n)}$ we have:

$$\|\zeta - \mu\|^2 = \sum_{j=1}^{d}(\zeta_j - \mu_j)^2 \leq \sum_{j=1}^{d} 4\sigma_j^2 \leq 4\sigma^2.$$

■

Appendix B. Quadratic Corner Saddle Point Case

We introduce the key ideas of the analysis in a simplified setting when: 1) the function f is quadratic, 2) the gradient is small, 3) the current iterate is located in a corner of the constraint space. Intuitively, this represents the key challenge of the constrained saddle escape problem since its NP-hardness comes from the hardness of the matrix copositivity problem in Remark 6 (i.e. the function is exactly quadratic and has no gradient at the current iterate which lies in the intersection of all constraints). We refer to this setting as the Quadratic Corner Saddle Point problem defined formally below. By shifting the coordinate system, w.l.o.g. we can assume that the saddle point is $\mathbf{0}$ and $f(\mathbf{0}) = 0$[3]. If $\mathbf{0}$ is a (δ, r)-QCSP (as defined below), the objective can be decreased by δ within a ball of radius r.

Definition 12 (Quadratic Corner Saddle Point) *Let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{0}\}$. For $\delta, r > 0$ and function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$, we say that a point $\mathbf{0}$ is a:*

- *(δ, r)-Quadratic Corner Saddle Point ((δ, r)-QCSP) if $\min_{\mathbf{x} \in \mathcal{B}(r) \cap S} f(\mathbf{x}) < -\delta$.*
- *(δ, r)-boundary QCSP if $\min_{\mathbf{x} \in \mathcal{B}(r) \cap \partial S} f(\mathbf{x}) < -\delta$.*

Algorithm 1: HOUDINIESCAPECORNER: Escaping from a corner for a quadratic function

input : Starting point $\mathbf{x}$, feasible set $S = \{\mathbf{y} \mid \mathbf{A}(\mathbf{y} - \mathbf{x}) \le \mathbf{0} \in \mathbb{R}^k\}$
parameters: δ and r from definition of (δ, r)-QCSP **for** $\mathcal{I} \in 2^{[k]}$ – *every subset of constraints* **do**

  // Optimize in $\mathcal{A}$
  $\mathcal{A} \leftarrow \{\mathbf{y} \mid \mathbf{A}_i^\top(\mathbf{y} - \mathbf{x}) = 0 \text{ for } i \in \mathcal{I}\}$, where $\mathbf{A}_i$ is the i-th row of $\mathbf{A}$
  $\mathbf{y} \leftarrow$ FINDINSIDECORNER$(\mathbf{x}, \mathcal{A})$
  if $\mathbf{y} \in S$ *and* $f(\mathbf{y}) < f(\mathbf{x}) - \frac{\delta}{2}$ **then**
    return $\mathbf{y}$

return $\perp$

Algorithm 2: FINDINSIDECORNER$(\mathbf{x}, \mathcal{A})$

input : Corner $\mathbf{x}$, affine subspace $\mathcal{A}$ with $\mathbf{x} \in \mathcal{A}$
parameters: δ and r from Definition 12, step size $\eta = \frac{1}{L}$, number of iterations $T = \tilde{O}(\frac{Lr^2}{\delta})$

Sample $\xi \sim \mathcal{N}(\mathbf{0}, I)$, $\mathbf{x}_0 \leftarrow \mathrm{Proj}_{\mathcal{A}}(\mathbf{x} + \xi)$
for $t = 0, \ldots, T-1$ **do**
  // Power method step
  $\mathbf{x}_{t+1} \leftarrow \mathrm{Proj}_{\mathcal{A}}(\mathbf{x}_t - \eta(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x})))$
$\mathbf{e} \leftarrow r\frac{\mathbf{x}_T - \mathbf{x}}{\|\mathbf{x}_T - \mathbf{x}\|}$
return $\mathbf{x} + \mathbf{e}$

In this section, we show how to escape from a (δ, r)-QCSP, namely we prove the following Theorem:

Theorem 13 (Theorem 7 restated) *Let $\delta, r > 0$. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ with $\lambda_{|max|}(\mathbf{M}) \le L$ and let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{0}\}$ be defined by k linear inequalities. If $\mathbf{0}$ is a (δ, r)-QCSP, then Algorithm 1 with*

3. Algorithms 1 and 2 don't require saddle point $\mathbf{x}$ to be $\mathbf{0}$. All the statements are trivially adapted for the case when $\mathbf{x}$ is not $\mathbf{0}$

probability at least $1-\xi$ *finds a point* $\mathbf{x} \in S \cap \mathcal{B}(r)$ *with* $f(\mathbf{x}) < -\Omega(\delta)$ *using* $O\left(\frac{Lr^2k2^k}{\delta}\log\frac{1}{\xi}\right)$ *deterministic gradient oracle calls.*

For the rest of the section, we assume that $\mathbf{0}$ is a (δ, r)-QCSP, i.e. $\min_{\mathbf{x}\in S\cap\mathcal{B}(r)} f(\mathbf{x}) < -\delta$. We consider two cases depending on whether $\mathbf{0}$ is a (δ, r)-boundary QCSP.

Case 1: 0 is a (δ, r)-boundary QCSP. For a subset of inequality constraints $\mathcal{I} \subseteq [k]$ we define the subspace where these constraints are active: $\mathcal{A}_{\mathcal{I}} = \{\mathbf{x} \mid \mathbf{A}_i^\top \mathbf{x} = 0 \text{ for all } i \in \mathcal{I}\}$. Let $\mathcal{I}$ be a maximal[4] subset of constraints such that $\min_{\mathbf{x}\in\mathcal{A}_{\mathcal{I}}\cap\mathcal{B}(r)} f(\mathbf{x}) < -\delta$. If $\mathbf{P}$ is a projection operator on $\mathcal{A}_{\mathcal{I}}$, it suffices to optimize $g(\mathbf{x}) := f(\mathbf{P}\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\mathbf{PMP})\mathbf{x}$. Therefore, we reduced the original problem to minimizing a different quadratic form in the same feasible set. For any $i \in \mathcal{I}$, $\mathbf{A}_i^\top \mathbf{P}\mathbf{x} \le 0$ holds trivially, since $\mathbf{A}_i^\top \mathbf{y} = 0$ for any $\mathbf{y} \in \mathcal{A}$, and hence constraints from $\mathcal{I}$ can be ignored. If a constraint not from $\mathcal{I}$ is active in $\mathbf{P}\mathbf{x}$, then $f(\mathbf{P}\mathbf{x}) \ge -\delta$, since $\mathcal{I}$ is a maximal subset of constraints with $\min_{\mathbf{x}\in\mathcal{A}_{\mathcal{I}}\cap\mathcal{B}(r)} < -\delta$. Therefore, this reduces Case 1 to Case 2.

Case 2: 0 is not a (δ, r)-boundary QCSP. In this case, we show that any $\mathbf{x} \in \mathcal{B}(r)$ with $f(\mathbf{x}) < -\delta$ must lie in S, and for $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x}$ it suffices to find the eigenvector corresponding to the smallest eigenvalue of $\mathbf{M}$. We first show that there exists an eigenvector that improves the objective.

Lemma 14 *Let* $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x}$ *and* S *be a closed linear cone. If* $\mathbf{0}$ *is not a* (δ, r)*-boundary QCSP for* $\delta, r > 0$*, then the following statements are equivalent:*

1. $\mathbf{0}$ *is* (δ, r)*-QCSP, i.e.* $\min_{\mathbf{x}\in S\cap\mathcal{B}(r)} f(\mathbf{x}) < -\delta$.
2. *There exists an eigenvector* $\mathbf{e}$ *of* $\mathbf{M}$ *such that* $\mathbf{e} \in \operatorname{Int} S \cap \partial\mathcal{B}(r)$ *and* $f(\mathbf{e}) < -\delta$.

Proof
2) $\implies$ 1) follows trivially by definition of (δ, r)-QCSP.
1) $\implies$ 2). We show that minimizer $\mathbf{x}^\star$ lies in $\partial(\mathcal{B}(r))$, and, since it's a local minimum, by the method of Lagrangian multipliers, we show $\nabla f(\mathbf{x}^\star) = \lambda\mathbf{x}^\star$ for some λ, where $\nabla f(\mathbf{x}^\star) = \mathbf{M}\mathbf{x}^\star$.

Let $\mathbf{x}^\star$ be the minimizer of f on $S \cap \mathcal{B}(r)$. Then $\mathbf{x}^\star \in \partial\mathcal{B}(r)$, since otherwise, we can rescale the minimizer to be on the boundary. I.e. for $\mathbf{y} = \frac{r}{\|\mathbf{x}^\star\|}\mathbf{x}^\star$ we have $f(\mathbf{y}) = \frac{r^2}{\|\mathbf{x}^\star\|^2} f(\mathbf{x}^\star) < f(\mathbf{x}^\star)$. Furthermore, $\mathbf{y} \in S$ since $\mathbf{x}^\star \in S$ and S is a linear cone, and $\mathbf{y} \in \partial\mathcal{B}(r)$ since $\|\mathbf{y}\| = r$.

From the above reasoning we have that $\mathbf{x}^\star \in S \cap \partial\mathcal{B}(r)$, and by the assumption of the lemma we have $f(\mathbf{x}^\star) < -\delta$. Since

$$\min_{\mathbf{x}\in\partial S\cap\partial\mathcal{B}(r)} f(\mathbf{x}) \ge \min_{\mathbf{x}\in\partial S\cap\mathcal{B}(r)} f(\mathbf{x}) \ge -\delta,$$

the minimizer $\mathbf{x}^\star$ lies in $\operatorname{Int} S \cap \partial\mathcal{B}(r)$. Since $\mathbf{x}^\star \in \partial\mathcal{B}(r)$, it satisfies the constraint $g(\mathbf{x}^\star) = \|\mathbf{x}^\star\|^2 - r^2 = 0$, and by the method of Lagrangian multipliers we must have $\nabla f(\mathbf{x}^\star) = \lambda\nabla g(\mathbf{x}^\star)$ or equivalently $\mathbf{M}\mathbf{x}^\star = \lambda\mathbf{x}^\star$ for some λ. Therefore, $\mathbf{x}^\star$ is an eigenvector of $\mathbf{M}$ and $f(\mathbf{x}^\star) < -\delta$. ■

Finding an exact eigenvector might be impossible. Hence, we show that finding $\mathbf{x} \in \mathcal{B}(r)$ with $f(\mathbf{x}) < -\delta$ suffices, since either $\mathbf{x}$ or $-\mathbf{x}$ are in S.

Lemma 15 *Let* $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x}$ *and* S *be a closed linear cone. For* $\delta, r > 0$ *and* $\hat{\mathbf{x}} \in \partial\mathcal{B}(r)$*, if the following conditions hold, then either* $\hat{\mathbf{x}} \in S$ *or* $-\hat{\mathbf{x}} \in S$*:*

4. As we don't know $\mathcal{I}$, Algorithm 1 tries all subsets of constraints.

1. *$\mathbf{0}$ is a (δ, r)-QCSP, i.e.* $\min_{\mathbf{x}\in S\cap\mathcal{B}(r)} f(\mathbf{x}) < -\delta$,
2. *$\mathbf{0}$ is not a (δ, r)-boundary QCSP, i.e.* $\min_{\mathbf{x}\in \partial S\cap\mathcal{B}(r)} f(\mathbf{x}) \geq -\delta$,
3. *$f(\hat{\mathbf{x}}) < -\delta$.*

Proof We know that there exists an eigenvector $\mathbf{e} \in S$ with $f(\mathbf{e}) < -\delta$. If both $\hat{\mathbf{x}}$ and $-\hat{\mathbf{x}}$ don't belong to S, we consider an arc on $\partial\mathcal{B}(r)$ connecting $\mathbf{e}$ and $\hat{\mathbf{x}}$ (or $-\hat{\mathbf{x}}$, see below). We will show that for any point $\mathbf{x}$ on the arc, $f(\mathbf{x}) < -\delta$, and, since the arc intersects with ∂S, this contradicts assumption that $\mathbf{0}$ is not a (δ, r)-boundary QCSP.

For contradiction, assume that both $-\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}$ don't belong to S. From Lemma 14 we know that there exists an eigenvector $\mathbf{e}$ with $f(\mathbf{e}) < -\delta$ and $\mathbf{e} \in S \cap \partial\mathcal{B}(r)$. Either $\hat{\mathbf{x}}^\top\mathbf{e} \geq 0$ or $-\hat{\mathbf{x}}^\top\mathbf{e} \geq 0$, and w.l.o.g. we assume $\hat{\mathbf{x}}^\top\mathbf{e} \geq 0$. Consider an arc between $\hat{\mathbf{x}}$ and $\mathbf{e}$ on $\partial\mathcal{B}(r)$:

$$U = \{\alpha\hat{\mathbf{x}} + \beta\mathbf{e} \mid \alpha, \beta \geq 0\} \cap \partial\mathcal{B}(r).$$

Since U is a connected set, $\hat{\mathbf{x}} \notin S$ and $\mathbf{e} \in S$, there exists $\mathbf{x} \in U \cap \partial S$. Since $\mathbf{x} = \alpha\hat{\mathbf{x}} + \beta\mathbf{e}$ for some $\alpha, \beta \geq 0$, we have:

$$\begin{aligned}\frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x} &= \frac{1}{2}(\alpha\hat{\mathbf{x}} + \beta\mathbf{e})^\top\mathbf{M}(\alpha\hat{\mathbf{x}} + \beta\mathbf{e}) \\ &= \frac{1}{2}\alpha^2\hat{\mathbf{x}}^\top\mathbf{M}\hat{\mathbf{x}} + \alpha\beta\mathbf{e}^\top\mathbf{M}\hat{\mathbf{x}} + \frac{1}{2}\beta^2\mathbf{e}^\top\mathbf{M}\mathbf{e}.\end{aligned}$$

By our assumption, $\frac{1}{2}\hat{\mathbf{x}}^\top\mathbf{M}\hat{\mathbf{x}} < -\delta$ and $\frac{1}{2}\mathbf{e}^\top\mathbf{M}\mathbf{e} < -\delta$. Since $\hat{\mathbf{x}}^\top\mathbf{e} \geq 0$ and $\mathbf{e}$ is an eigenvector of $\mathbf{e}$ with eigenvalue at most $\frac{f(\mathbf{e})}{\|\mathbf{e}\|^2/2} = -\frac{2\delta}{r^2}$, we have

$$\alpha\beta\mathbf{e}^\top\mathbf{M}\hat{\mathbf{x}} < -\frac{2\delta}{r^2}\alpha\beta\mathbf{e}^\top\hat{\mathbf{x}},$$

where we used that all terms are non-negative. Finally, using $\|\hat{\mathbf{x}}\| = \|\mathbf{e}\| = \|\alpha\hat{\mathbf{x}} + \beta\mathbf{e}\| = r$:

$$\begin{aligned}\frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x} &< -\frac{\delta}{r^2}(\alpha^2\|\hat{\mathbf{x}}\|^2 + 2\alpha\beta\mathbf{e}^\top\hat{\mathbf{x}} + \beta^2\|\mathbf{e}\|^2) \\ &= -\frac{\delta}{r^2}\|\alpha\hat{\mathbf{x}} + \beta\mathbf{e}\|^2 \\ &= -\delta,\end{aligned}$$

contradicting our assumption that $\mathbf{0}$ is not a (δ, r)-boundary QCSP. ■

Finding $\mathbf{x} \in \mathcal{B}(r)$ with $f(\mathbf{x}) < -\delta$ might be difficult if $\min_{\mathbf{x}\in\mathcal{B}(r)} f(\mathbf{x})$ is very close to $-\delta$. Instead we aim to find $\mathbf{x} \in \mathcal{B}(r)$ with $f(\mathbf{x}) < -(1-\varepsilon)\delta$ for a constant $\varepsilon \in (0, 1)$. As we show next, such a point can be found in polynomial time using the power method applied to the matrix $I - \frac{\mathbf{M}}{L}$. Since $\lambda_{|max|}(\mathbf{M}) \leq L$, all eigenvalues of $I - \frac{\mathbf{M}}{L}$ are between 0 and 2. Since there exists an eigenvector $\mathbf{e} \in \partial\mathcal{B}(r)$ of $\mathbf{M}$ with $\frac{1}{2}\mathbf{e}^\top\mathbf{M}\mathbf{e} < -\delta$, we have $\lambda_{\min}(\mathbf{M}) < \frac{-2\delta}{\|\mathbf{e}\|^2} = \frac{-2\delta}{r^2}$, and hence the largest eigenvalue of $I - \frac{\mathbf{M}}{L}$ is at least $1 - \frac{\lambda_{\min}(\mathbf{M})}{L} \geq 1 + \frac{2\delta}{Lr^2}$. Finding $\mathbf{x}$ with $\frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x} < -(1-\varepsilon)\delta$ is equivalent to finding $\mathbf{x}$ with $\mathbf{x}^\top(I - \frac{\mathbf{M}}{L})\mathbf{x} \geq (1 + \frac{2(1-\varepsilon)\delta}{Lr^2})\|\mathbf{x}\|^2$, and the power method achieves this in $O(\log d + \frac{Lr^2}{\varepsilon\delta})$ iterations (see Lemma 16). Finally, we show that Algorithm 2 emulates the power method on matrix $I - \frac{\mathbf{M}}{L}$.

Lemma 16 *Let* $\delta, r > 0$, $\mathbf{x} \in \mathbb{R}^f$ *and* $\varepsilon \in (0,1)$. *Let* $f(\mathbf{y}) = \frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \mathbf{M}(\mathbf{y}-\mathbf{x})$ *with* $\lambda_{|max|}(\mathbf{M}) \le L$. *Let* $\mathcal{A}$ *be an affine subspace of* $\mathbb{R}^d$ *such that* $\mathbf{x} \in \mathcal{A}$. *If* $\min_{\mathbf{y} \in \mathcal{A} \cap \mathcal{B}(\mathbf{x}, r)} f(\mathbf{y}) < -\delta$, *then Algorithm 2 with* $\eta = \frac{1}{L}$ *with probability* $1 - O(\xi)$ *finds* $\mathbf{y} \in \mathcal{A} \cap \partial\mathcal{B}(\mathbf{x}, r)$ *with* $f(\mathbf{y}) \le -(1-\varepsilon)\delta$ *after* $T = O\left(\frac{Lr^2}{\varepsilon\delta}\log\left(\frac{Lrd}{\xi\varepsilon\delta}\right)\right)$ *iterations*[5].

Proof We show that Algorithm 2 performs a power iteration on matrix $I - \eta\mathbf{PMP}$, where $\mathbf{P}$ is the projection operator on linear subspace corresponding to $\mathcal{A}$. We show that $\lambda_{\min}(\mathbf{PMP})$ corresponds to $\lambda_{|max|}(I - \eta\mathbf{PMP})$. Finally, we use Fact 10 to establish the convergence rate.

By shifting the coordinate system so that $\mathbf{x}$ becomes $\mathbf{0}$, we instead optimize function $g(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top\mathbf{M}\mathbf{y}$ on set $\mathcal{B}(r) \cap \mathcal{A}_0$, where $\mathcal{A}_0 = \{\mathbf{y} - \mathbf{x} \mid \mathbf{y} \in \mathcal{A}\}$ is the linear subspace parallel to $\mathcal{A}$. Let $\mathbf{P} \in \mathbb{R}^{d \times d}$ be the projection operator on $\mathcal{A}_0$. Then for any $\mathbf{y} \in \mathcal{A}_0$ we have $\mathbf{y} = \mathbf{Py}$ and for any $\mathbf{y} \in \mathbb{R}^d$ we have $\operatorname{Proj}_{\mathcal{A}_0}(\mathbf{y}) = \mathbf{Py}$. Defining $\mathbf{y}_t = \mathbf{x}_t - \mathbf{x}$, we have:

$$\begin{aligned}
\mathbf{y}_{t+1} &= \mathbf{x}_{t+1} - \mathbf{x} \\
&= \operatorname*{Proj}_{\mathcal{A}}(\mathbf{x}_t - \eta(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}))) - \mathbf{x} && \text{(Algorithm 2, Line 4)} \\
&= \operatorname*{Proj}_{\mathcal{A}_0}(\mathbf{x}_t - \mathbf{x} - \eta(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}))) && (\operatorname*{Proj}_{\mathcal{A}}(\mathbf{y}) - \mathbf{x} = \operatorname*{Proj}_{\mathcal{A}_0}(\mathbf{y} - \mathbf{x})) \\
&= \mathbf{P}(\mathbf{x}_t - \mathbf{x} - \eta\mathbf{M}(\mathbf{x}_t - \mathbf{x})) && (\nabla f(\mathbf{x}) = \mathbf{Mx} \text{ and } \operatorname*{Proj}_{\mathcal{A}_0}(\mathbf{y}) = \mathbf{Py}) \\
&= \mathbf{y}_t - \eta\mathbf{PM}\mathbf{y}_t && \text{(Definition of } \mathbf{y}_t) \\
&= \mathbf{y}_t - \eta\mathbf{PMP}\mathbf{y}_t && (\mathbf{P}\mathbf{y}_t = \mathbf{y}_t \text{ since, by induction, } \mathbf{y}_t \in \mathcal{A}_0) \\
&= (I - \eta\mathbf{PMP})^{t+1}\mathbf{y}_0.
\end{aligned}$$

Therefore, Algorithm 2 performs a power iteration for matrix $\mathbf{B} = I - \eta\mathbf{PMP}$. Since $\lambda_{|max|}(\mathbf{PMP}) \le \lambda_{|max|}(\mathbf{M}) \le L$ and $\eta = \frac{1}{L}$, all eigenvalues of $\mathbf{B}$ lie in $[0, 2]$, and its largest eigenvalue is $1 - \eta\lambda_{\min}(\mathbf{PMP})$.

Since there exists $\mathbf{e}$ with $\|\mathbf{e}\| = r$ and $h(\mathbf{e}) < -\delta$, we have

$$\lambda_{\min}(\mathbf{PMP}) \le -\frac{f(\mathbf{e})}{\|\mathbf{e}\|^2/2} < -\frac{2\delta}{r^2},$$

which implies $\lambda_{|max|}(\mathbf{B}) = 1 + \frac{2\delta}{Lr^2}$. Our goal is to find $\mathbf{y} \in \partial\mathcal{B}(r)$ with $f(\mathbf{y}) < -(1-\varepsilon)\delta$, meaning

$$\frac{\mathbf{y}^\top(\mathbf{PMP})\mathbf{y}}{\|\mathbf{y}\|^2} < -(1-\varepsilon)\frac{2\delta}{r^2},$$

which is equivalent to

$$\frac{\mathbf{y}^\top\mathbf{B}\mathbf{y}}{\|\mathbf{y}\|^2} > 1 + (1-\varepsilon)\frac{2\delta}{Lr^2}.$$

5. This is the only statement in this section where we consider a non-zero saddle point, due to its non-trivial role in Algorithm 2

By Fact 10, with probability $1 - O(\xi)$ we find such $\mathbf{y}$ after the following number of iterations:

$$T = O\left(\frac{\log d - \log(\xi\varepsilon\frac{2\delta}{Lr^2})}{\log\frac{1+2\delta/Lr^2}{1+(1-\varepsilon)\cdot 2\delta/Lr^2}}\right)$$
$$= O\left(\frac{\log\left(\frac{Lrd}{\xi\varepsilon\delta}\right)}{\log\left(1+\frac{\varepsilon\delta}{Lr^2}\right)}\right)$$
$$= O\left(\frac{Lr^2}{\varepsilon\delta}\log\left(\frac{Lrd}{\xi\varepsilon\delta}\right)\right).$$

Rescaling $\mathbf{y}$ so that $\|\mathbf{y}\| = r$ finishes the proof. ■

We now prove Theorem 7. First, by exhaustive search we guess a maximal subset of active constraints $\mathcal{I}$ such that subspace $\mathcal{A}_\mathcal{I}$ formed by these linear constraints has $\mathbf{x} \in \mathcal{B}(r) \cap S$ with $f(\mathbf{x}) < \delta$. Using Algorithm 2, we find $\mathbf{y} \in \mathcal{B}(r) \cap \mathcal{A}_\mathcal{I}$ with $f(\mathbf{y}) < -(1-\varepsilon)\delta$. Then $\mathbf{y} \in S$ by Lemma 16, since $\mathcal{I}$ is a maximal subset of constraints with an escape direction.

Definition 17 *A set S is a* linear cone *if $\mathbf{x} \in S$ implies $\alpha\mathbf{x} \in S$ for all $\alpha \geq 0$.*

If S is a linear cone, then scaling a vector preserves its belonging to S. In particular, $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq 0\}$ is a linear cone.

Theorem 18 (Theorem 7 restated) *Let $\delta, r > 0$. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x}$ with $\lambda_{|max|}(\mathbf{M}) \leq L$ and let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$ be defined by k linear inequality constraints. If $\mathbf{x}$ is a (δ, r)-QCSP, then Algorithm 1 with probability at least $1-\xi$ finds a point $\mathbf{x} \in S \cap \mathcal{B}(r)$ with $f(\mathbf{x}) < -\Omega(\delta)$ using $\tilde{O}\left(\frac{Lr^2k2^k}{\delta}\right)$ deterministic gradient oracle calls.*

Proof

For $\mathcal{I} \subseteq [k]$, we define $\mathcal{A}_\mathcal{I} = \{\mathbf{x} \mid \mathbf{A}_i^\top\mathbf{x} = 0, i \in \mathcal{I}\}$. By induction, we'll prove the following: for any $\mathcal{I} \in [k]$, if Algorithm 1 didn't find point $\mathbf{x}$ with $f(\mathbf{x}) < -(1 - \frac{|\mathcal{I}|+1}{k}\varepsilon)\delta$ after executing Algorithm 2 on all sets $\mathcal{A}_{\tilde{\mathcal{I}}}$ such that $\mathcal{I} \subseteq \tilde{\mathcal{I}} \subseteq [k]$, then

$$\min_{\mathbf{x}\in\mathcal{A}_\mathcal{I}\cap S\cap\mathcal{B}(r)} f(\mathbf{x}) \geq -\left(1 - \frac{|\mathcal{I}|}{k}\varepsilon\right)\delta.$$

Basically, we have a subroutine which can return $1-\varepsilon$ approximation, and, since this approximation factor may accumulate with each additional active constraint, we instead use approximation $1 - \frac{\varepsilon}{k}$ for every set of constraints.

We fix $\mathcal{I}$ and assume that the statement holds for all $\tilde{\mathcal{I}}$ such that $\mathcal{I} \subsetneq \tilde{\mathcal{I}} \subseteq [k]$. By induction hypothesis, if for all such $\tilde{\mathcal{I}}$ we didn't find point $\mathbf{x}$ with $f(\mathbf{x}) < -(1 - \frac{|\mathcal{I}|+2}{k}\varepsilon)\delta$, then

$$\min_{\mathbf{x}\in\mathcal{A}_{\tilde{\mathcal{I}}}\cap S\cap\mathcal{B}(r)} f(\mathbf{x}) \geq -(1 - \frac{|\mathcal{I}|+1}{k}\varepsilon)\delta$$

for all such $\tilde{\mathcal{I}}$. If

$$\min_{\mathbf{x}\in\mathcal{A}_\mathcal{I}\cap S\cap\mathcal{B}(r)} f(\mathbf{x}) < -(1 - \frac{|\mathcal{I}|}{k}\varepsilon)\delta,$$

then, since on the boundary we have

$$\min_{\mathcal{I} \subsetneq \tilde{\mathcal{I}} \subseteq [k]} \min_{\mathbf{x} \in \mathcal{A}_{\tilde{\mathcal{I}}} \cap S \cap \mathcal{B}(r)} f(\mathbf{x}) \geq -(1 - \frac{|\mathcal{I}|+1}{k}\varepsilon)\delta,$$

by Lemma 15 and Lemma 16 we can find $\mathbf{x} \in \mathcal{A}_{\mathcal{I}} \cap S \cap \mathcal{B}(r)$ with $f(\mathbf{x}) < -(1 - \frac{|\mathcal{I}|+1}{k}\varepsilon)\delta$ using $\tilde{O}\left(\frac{Lkr^2}{\varepsilon\delta}\right)$ gradient computations. Hence, if the algorithm didn't find such $\mathbf{x}$, then, as required:

$$\min_{\mathbf{x} \in \mathcal{A}_{\mathcal{I}} \cap S \cap \mathcal{B}(r)} f(\mathbf{x}) \geq -(1 - \frac{|\mathcal{I}|}{k}\varepsilon)\delta.$$

Since there 2^k possible subsets of constraints, the total number of gradient oracle calls is $\tilde{O}\left(\frac{Lkr^2 2^k}{\varepsilon\delta}\right)$.
■

Appendix C. Algorithm and Proofs for General Case

In this section, we present Algorithms 3 and 4 and show that they can escape from a saddle point. Algorithm 3 computes a quadratic approximation of the function in all affine subspaces, defined by all possible intersections of active constraints. For each affine subspace, it calls Algorithm 4 as a subroutine, which tries to find an escape direction in this subspace.

C.1. Algorithm 3: Quadratic approximation

We first focus on the Lines 1-8 of Algorithm 3. The goal is to build a quadratic approximation of the objective with some additional properties (see below). To simplify the presentation, w.l.o.g. by shifting the coordinate system, we assume that the current saddle point is $\mathbf{0}$ and consider the quadratic approximation of the function: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x} + \mathbf{x}^\top \nabla f(\mathbf{0})$. Since f is ρ-Lipschitz and $r = \sqrt[3]{\delta/\rho}$, in $\mathcal{B}(r)$ the quadratic approximation deviates from f by at most $\frac{\delta}{6}$ (see derivation before Definition 4). For a small value ξ (to be specified later), we instead analyze a noisy function $f'(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x} + \mathbf{x}^\top \mathbf{v}$, where (all lines refer to Algorithm 3):

1. $\mathbf{M} = \mathbf{H} + \gamma I$ (Line 8), where $\mathbf{H}$ is an approximation of the Hessian (Line 5) and γ is small uniform noise, which guarantees that $\mathbf{M}$ is non-degenerate with probability 1.
2. $\mathbf{v} = \mathbf{g} + \zeta$ (Line 7), where $\mathbf{g}$ is an approximation of the gradient and ζ is small Gaussian noise, which guarantees that all coordinates of $\mathbf{v}$ are sufficiently separated from 0 w.h.p. Additionally, with probability 1, linear systems of the form $(\mathbf{M} - \mu I)\mathbf{x} = \mathbf{v}$ with $\mathrm{rank}(\mathbf{M} - \mu I) < d$ don't have any solutions, simplifying the analysis.[6]

We show that f' is a good approximation of f, and first, we need to show that $\mathbf{H}$ and $\mathbf{g}$ are good approximations of the Hessian and the gradient respectively. For the gradient, we use algorithm VRSG from Line 1, which w.h.p. estimates the gradient with precision $\tilde{\sigma}$ using $\tilde{O}(\frac{\sigma^2}{\tilde{\sigma}^2})$ stochastic gradient oracle calls (see Appendix C, Lemma 20). For the Hessian, since the i-th column of $\nabla^2 f(\mathbf{0})$ is by definition $\lim_{\tau \to 0} \frac{\nabla f(\tau \mathbf{e}_i) - \nabla f(\mathbf{0})}{\tau}$, using a sufficiently small τ and approximating $\nabla f(\tau \mathbf{e}_i)$ and $\nabla f(\mathbf{0})$ using VRSG, we find good approximation of $\nabla^2 f(\mathbf{x})$. By selecting parameters as in Algorithm 3 and by guaranteeing that noises ζ and γ are bounded, we show the following result:

6. This also holds in the subspaces, see proof of Theorem 26).

Algorithm 3: HOUDINIESCAPE$(\mathbf{x}, S, \delta)$: Escaping from a saddle point

input : Saddle point $\mathbf{x}$, feasible set $S = \{\mathbf{y} \mid \mathbf{A}\mathbf{y} \le \mathbf{b} \in \mathbb{R}^k\}$, δ from definition of δ-SOSP
output : $\mathbf{u} \in S \cap \mathcal{B}(\mathbf{x}, r)$ with $f(\mathbf{u}) < f(\mathbf{x}) - \frac{\delta}{3}$, or reports that $\mathbf{x}$ is a δ-SOSP
parameters: $r = \sqrt[3]{\delta/\rho}$ (Definition 4), noise parameter ξ

Let $\mathrm{VRSG}(\mathbf{x}, \tilde{\sigma})$ be an algorithm that returns $\mathbf{g}$ such that $\|\mathbf{g}^{(0)} - \nabla f(\mathbf{x})\| \le \tilde{\sigma}$ w.h.p.
// Compute $\mathbf{H}$ - approximate Hessian
$\theta \leftarrow \Theta(\frac{r}{\sqrt{d}}), \quad \tilde{\sigma} \leftarrow \Theta(\frac{\rho r^2}{d})$
$\mathbf{g}^{(0)} \leftarrow \mathrm{VRSG}(\mathbf{x}, \tilde{\sigma})$
for $\mathbf{e}_1, \dots, \mathbf{e}_d$ *– the standard basis in* $\mathbb{R}^d$ **do** $\mathbf{g}_i \leftarrow \mathrm{VRSG}(\mathbf{x} + \theta \mathbf{e}_i, \tilde{\sigma})$;
Let $\mathbf{H} \in \mathbb{R}^{d \times d}$ whose i-th column is $\frac{\mathbf{g}_i - \mathbf{g}^{(0)}}{\theta}$

// $f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \mathbf{h}^\top \mathbf{v} + \frac{1}{2}\mathbf{h}^\top \mathbf{M}\mathbf{h}$
Sample $\zeta \sim \mathcal{N}(\mathbf{0}, \xi \frac{\delta}{r\sqrt{d}} \cdot I), \quad \gamma \sim U[-\xi, \xi]$
$\mathbf{v} \leftarrow \mathrm{VRSG}(\mathbf{x}, \Theta(\delta/r)) + \zeta$
$\mathbf{M} \leftarrow \mathbf{H} + \gamma I$

for $\mathcal{I}$ *– every subset of constraints* **do**
  $\mathcal{A} \leftarrow \{\mathbf{y} \mid \mathbf{A}_i^\top \mathbf{x} = b_i,\ i \in \mathcal{I}\}$, where $\mathbf{A}_i$ is the i-th row of $\mathbf{A}$ // Optimize in $\mathcal{A}$
  $\mathbf{p} \leftarrow \mathrm{Proj}_{\mathcal{A}}(\mathbf{x})$ // Ball center in $\mathcal{A}$
  if $\|\mathbf{p} - \mathbf{x}\| > r$ **then continue**;

  // Optimization of $g(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top \mathbf{M}_\perp \mathbf{y} + \mathbf{y}^\top \mathbf{v}_\perp$ in $S_\perp \cap \mathcal{B}_{\dim \mathcal{A}}(r_\perp)$
  $r_\perp \leftarrow \sqrt{r^2 - \|\mathbf{p} - \mathbf{x}\|^2}$
  Let $\mathbf{O} \in \mathbb{R}^{d \times \dim \mathcal{A}}$ be an orthonormal basis of $\mathcal{A}$
  $S_\perp \leftarrow \{\mathbf{y} \mid \mathbf{A}(\mathbf{p} + \mathbf{O}\mathbf{y}) \le \mathbf{b}\}$ // $\mathbf{y} \in S_\perp$ guarantees that $\mathbf{p} + \mathbf{O}\mathbf{y} \in S$
  $\mathbf{M}_\perp \leftarrow \mathbf{O}^\top \mathbf{M}\mathbf{O}$
  $\mathbf{v}_\perp \leftarrow \mathbf{O}^\top (\mathbf{v} + \mathbf{M}(\mathbf{p} - \mathbf{x}))$

  $\mathbf{u} \leftarrow$ FINDINSIDE$(\mathbf{x}, \delta, \mathbf{p}, \mathbf{O}, (\mathbf{M}_\perp, \mathbf{v}_\perp), (r_\perp, S_\perp))$ // Algorithm 4
  if $\mathbf{u} \ne \perp$ **then return** $\mathbf{u}$;
report that $\mathbf{x}$ is a δ-SOSP

Algorithm 4:
FINDINSIDE$(\mathbf{x}, \delta, \mathbf{p}, \mathbf{O}, (\mathbf{M}_\perp, \mathbf{v}_\perp), (r_\perp, S_\perp))$

input : $\mathbf{x}$ – saddle point, δ from definition of δ-SOSP, $\mathbf{p}$ – projection of $\mathbf{x}$ on $\mathcal{A}$ (affine subspace corresponding to active constraints), $\mathbf{O}$ – orthonormal basis of $\mathcal{A}$,
$g(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top \mathbf{M}_\perp \mathbf{y} + \mathbf{y}^\top \mathbf{v}_\perp$ – objective in $\mathbb{R}^{\dim \mathcal{A}}$, $S_\perp \cap \mathcal{B}(r_\perp)$ – feasible set in $\mathbb{R}^{\dim \mathcal{A}}$

output : $\mathbf{p} + \mathbf{O}\mathbf{y}$ such that $\mathbf{y} \in S_\perp \cap \mathcal{B}(r_\perp)$ and $f(\mathbf{p} + \mathbf{O}\mathbf{y}) < f(\mathbf{x}) - \Omega(\delta)$, if it exists.

// **Case 1:** Large gradient

1 Let $\mathbf{y} \leftarrow \operatorname{argmin}_{\mathbf{y} \in S_\perp \cap \mathcal{B}(r_\perp)} \mathbf{y}^\top \mathbf{v}_\perp$

2 **if** $f(\mathbf{p} + \mathbf{O}\mathbf{y}) < f(\mathbf{x}) - \frac{\delta}{3}$ **then return** $\mathbf{p} + \mathbf{O}\mathbf{y}$;

// **Case 2:** Solution is in $\operatorname{Int} \mathcal{B}(r_\perp)$

3 Let $\mathbf{y} \leftarrow -\mathbf{M}^{-1}\mathbf{v}$

4 **if** $\mathbf{y} \in S_\perp \cap \mathcal{B}(r_\perp)$ *and* $f(\mathbf{p} + \mathbf{O}\mathbf{y}) < f(\mathbf{x}) - \frac{\delta}{3}$ **then return** $\mathbf{p} + \mathbf{O}\mathbf{y}$;

// **Case 3:** Solution is in $\partial \mathcal{B}(r_\perp)$

// Matrix diagonalization

5 Let $\mathbf{Q}$ be an orthogonal matrix and $\mathbf{\Lambda} = diag(\lambda_1, \dots, \lambda_{\dim \mathcal{A}})$ be such that $\|\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} - \mathbf{M}_\perp\| \le \frac{\delta}{10r^2}$

// Rotation of the linear term

6 Let $\tilde{\mathbf{v}} \leftarrow \mathbf{Q}\mathbf{v}_\perp$ and $\tilde{v}_i$ be its i-th coordinate

// $\tilde{\mathbf{y}}$ from Line 9 must have norm r

7 Let $\{\mu\}_j$ be approximate real solutions of $\sum_{i=1}^d \frac{\tilde{v}_i^2}{(\mu - \lambda_i)^2} = r_\perp^2$

8 **for** *each* μ_j **do**

9   Let $\tilde{\mathbf{y}}$ be a vector such that its i-th coordinate is $\tilde{y}_i \leftarrow \frac{\tilde{v}_i}{\mu_j - \lambda_i}$

    // Correction of approximate solutions

10     $\mathbf{y} \leftarrow \operatorname{Proj}_{S_\perp \cap \mathcal{B}(r_\perp)}(\mathbf{Q}^\top \tilde{\mathbf{y}})$

11   **if** $f(\mathbf{p} + \mathbf{O}\mathbf{y}) < f(\mathbf{x}) - \frac{\delta}{3}$ **then return** $\mathbf{p} + \mathbf{O}\mathbf{y}$;

return $\perp$ // No solution in affine subspace $\mathcal{A}$

Lemma 19 *Let f satisfy Assumptions 1 and 2. Let $f'(\mathbf{x}) = f(\mathbf{0}) + \mathbf{x}^\top \mathbf{v} + \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$, where $\mathbf{v}$ and $\mathbf{M}$ are as in Algorithm 3. For $\delta > 0$, $r = \sqrt[3]{\delta/\rho}$ we have $\|f'(\mathbf{x}) - f(\mathbf{x})\| < \frac{\delta}{2}$ for all $\mathbf{x} \in \mathcal{B}(r)$ w.h.p.*

Reducing Case $\mathbf{x}^\star \in \partial S$ to Case $\mathbf{x}^\star \in \operatorname{Int} S$. We now elaborate on Lines 13-17 of Algorithm 3. The goal is, as in Section B, to reduce the case $\mathbf{x}^\star \in \partial S$ to the case $\mathbf{x}^\star \in \operatorname{Int} S$. If $\mathbf{x}^\star \in \partial S$, then there exist a non-empty set $\mathcal{I}$ of constraints active at $\mathbf{x}^\star$. Then we work in affine subspace $\mathcal{A}$ where constraints from $\mathcal{I}$ are active. We parameterize $\mathcal{A}$: if $\mathbf{p} = \operatorname{Proj}_{\mathcal{A}}(\mathbf{x})$ and $\mathbf{O} \in \mathbb{R}^{d \times \dim \mathcal{A}}$ is an orthonormal basis of $\mathcal{A}$, then any point in $\mathcal{A}$ can be represented as $\mathbf{p} + \mathbf{O}\mathbf{y}$ for $\mathbf{y} \in \mathbb{R}^{\dim \mathcal{A}}$. Defining $g(\mathbf{y}) = f'(\mathbf{p} + \mathbf{O}\mathbf{y})$, since

$$\begin{aligned} g(\mathbf{y}) &= {}^1\!/\!_2(\mathbf{p} + \mathbf{O}\mathbf{y})^\top \mathbf{M}(\mathbf{p} + \mathbf{O}\mathbf{y}) + (\mathbf{O}\mathbf{y})^\top \mathbf{p} \\ &= {}^1\!/\!_2 \cdot \mathbf{y}^\top (\mathbf{O}^\top \mathbf{M}\mathbf{O})\mathbf{y} + \mathbf{y}^\top (\mathbf{O}^\top \mathbf{M}\mathbf{p} + \mathbf{O}^\top \mathbf{v}) + const, \end{aligned}$$

minimizing f' in $\mathcal{A} \cap S \cap \mathcal{B}(r)$ is equivalent to minimizing g in $S_\perp \cap \mathcal{B}(r_\perp)$ (Lines 13-17), where:

1. $S_\perp$ is a set of points $\mathbf{y} \in \mathbb{R}^{\dim \mathcal{A}}$ such that $\mathbf{p} + \mathbf{O}\mathbf{y} \in S$, namely $S_\perp = \{\mathbf{y} \mid \mathbf{A}_i(\mathbf{p} + \mathbf{O}\mathbf{y}) \le b_i,\ i \notin \mathcal{I}\}$. Hence, $S_\perp$ is defined by linear inequalities, similarly to S.
2. $r_\perp$ is a radius such that condition $\mathbf{y} \in \mathcal{B}(r_\perp)$ is equivalent to $\mathbf{p} + \mathbf{O}\mathbf{y} \in \mathcal{B}(r)$. Since $\mathbf{p}$ is the projection of $\mathbf{0}$ on $\mathcal{A}$, we have $\mathbf{O}^\top \mathbf{p} = \mathbf{0}$, and hence $\|\mathbf{p} + \mathbf{O}\mathbf{y}\|^2 = \|\mathbf{p}\|^2 + \|\mathbf{O}\mathbf{y}\|^2$. Since $\mathbf{O}$ is an orthonormal basis of $\mathcal{A}$, $\|\mathbf{O}\mathbf{y}\| = \|\mathbf{y}\|$, and hence $r_\perp = \sqrt{r^2 - \|\mathbf{p}\|^2}$.

For $\mathbf{y}^\star$ such that $\mathbf{x}^\star = \mathbf{p} + \mathbf{O}\mathbf{y}^\star$, no constraints from $S_\perp$ are active, and hence $\mathbf{x}^\star \in \operatorname{Int} S_\perp$.

C.2. Algorithm 4: Escaping when $\mathbf{y}^\star \in \operatorname{Int} S_\perp$

In this section, we find minimizer $\mathbf{y}^\star$ of function $g(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top \mathbf{M}_\perp \mathbf{y} + \mathbf{y}^\top \mathbf{v}_\perp + C$ in $S_\perp \cap \mathcal{B}(r_\perp)$, while assuming that $\mathbf{y}^\star \in \operatorname{Int} S_\perp$. Since the solutions we find can be approximate, we have to guarantee that the objective is not too sensitive to the change of its argument. It suffices to guarantee that $\|\mathbf{v}\|$ is bounded, since for any $\mathbf{y} \in \mathcal{B}(r_\perp)$ and perturbation $\mathbf{h}$ there exists $\tau \in [0, 1]$ such that:

$$\begin{aligned} |g(\mathbf{y}) - g(\mathbf{y} + \mathbf{h})| &= |(\nabla g(\mathbf{y} + \tau \mathbf{h}))^\top \mathbf{h}| \\ &\le (\|\nabla g(\mathbf{0})\| + L\|\mathbf{y} + \tau \mathbf{h}\|)\|\mathbf{h}_\perp\| \\ &\le (\|\mathbf{v}\| + L(r_\perp + \|\mathbf{h}\|))\|\mathbf{h}\|, \end{aligned}$$

where we used that the objective is L-smooth and hence $\|\nabla g(\mathbf{y} + \tau \mathbf{h})\| \le \|\nabla g(\mathbf{0})\| + L\|\mathbf{y} + \tau \mathbf{h}\|$. We consider the situation when $\|\mathbf{v}_\perp\|$ is large as a separate case. Otherwise, for $\mathbf{y}^\star$, there are only two options: either $\mathbf{y}^\star \in \operatorname{Int} \mathcal{B}(r_\perp)$ or $\mathbf{y}^\star \in \partial \mathcal{B}(r_\perp)$. Algorithm 4 handles these cases, as well as the case when $\|\mathbf{v}\|$ is large, separately.

Case 1: $\|\mathbf{v}_\perp\|$ is large (Lines 1-2). If $\|\mathbf{v}_\perp\|$ is large and we can find $\mathbf{y}$ with small $\mathbf{y}^\top \mathbf{v}_\perp$, the linear term alone suffices to improve the objective. We show that, if such $\mathbf{y}$ doesn't exist, then $g(\mathbf{y}^\star)$ requires $\mathbf{y}^\star \in \partial S_\perp$, which contradicts that $\mathbf{y}^\star \in \operatorname{Int} S_\perp$. Below we assume that $\|\mathbf{v}_\perp\|$ is bounded.

Case 2: $\mathbf{y} \in \operatorname{Int} \mathcal{B}(r_\perp)$ (Lines 3-4). In this case, $\mathbf{y}^\star$ is an unconstrained critical point of g, and hence it must satisfy $\nabla g(\mathbf{y}) = \mathbf{0}$, implying $\mathbf{M}_\perp \mathbf{y} + \mathbf{v}_\perp = \mathbf{0}$ which gives the unique solution $\mathbf{y} = -\mathbf{M}_\perp^{-1} \mathbf{v}_\perp$. since $\mathbf{M}$ is a perturbed matrix, so is $\mathbf{M}_\perp$, and hence $\mathbf{M}_\perp$ is non-degenerate with probability 1. It remains to verify that $\mathbf{y} \in \mathcal{B}(r_\perp) \cap S_\perp$ and $\mathbf{y}$ decreases the objective by $\Omega(\delta)$.

Case 3: $\mathbf{y} \in \partial\mathcal{B}(r_\perp)$ (Lines 5-11). Since the only active constraint at $\mathbf{y}^\star$ is $c(\mathbf{y}) = \frac{1}{2}(\|\mathbf{y}\|^2 - r_\perp^2) = 0$, by the KKT conditions, any critical point must satisfy $\nabla g(\mathbf{y}) = \mu \nabla c(\mathbf{y})$ for some $\mu \in \mathbb{R}$, which is equivalent to $\mathbf{M}_\perp \mathbf{y} + \mathbf{v}_\perp = \mu \mathbf{y}$. Hence, for any fixed μ, $\mathbf{y}$ must be a solution of linear system $(\mathbf{M}_\perp - \mu I)\mathbf{y} = -\mathbf{v}_\perp$. When $\mathbf{M}_\perp - \mu I$ is degenerate (i.e. μ is an eigenvalue of $\mathbf{M}_\perp$), due to perturbation of $\mathbf{v}_\perp$, the system doesn't have any solution with probability 1. It leaves us with the case when $\mathbf{M}_\perp - \mu I$ is non-degenerate, when there exists a unique solution $\mathbf{y}(\mu) := -(\mathbf{M}_\perp - \mu I)^{-1}\mathbf{v}_\perp$.

Diagonalization. Dependence of $(\mathbf{M}_\perp - \mu I)^{-1}$ on μ is non-trivial, but for a diagonal $\mathbf{M}_\perp$, the inverse can be found explicitly. Hence, we perform diagonalization of $\mathbf{M}_\perp$: we find orthogonal $\mathbf{Q}$ and diagonal $\mathbf{\Lambda}$ such that $\|\mathbf{M}_\perp - \mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}\| < \varepsilon$ (Line 5) in time $O(d^3 \log {}^1\!/\!_\varepsilon)$ [30]. Setting $\varepsilon = O({}^\delta\!/\!_{r_\perp^2})$, we guarantee that the function changes by at most $O(\delta)$ in $\mathcal{B}(r_\perp)$. The function $\mathbf{y}(\mu) = -(\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} - \mu I)^{-1}\mathbf{v}_\perp$ can be written as $\mathbf{Q}\mathbf{y}(\mu) = -(\mathbf{\Lambda} - \mu I)\mathbf{Q}\mathbf{v}_\perp$, and hence we work with rotated vectors $\tilde{\mathbf{y}}(\mu) := \mathbf{Q}\mathbf{y}(\mu)$ and $\tilde{\mathbf{v}} := \mathbf{Q}\mathbf{v}_\perp$ (Line 6).

Finding candidate μ. Since $\tilde{\mathbf{y}}(\mu) = -(\mathbf{M}_\perp - \mu I)^{-1}\tilde{\mathbf{v}}$, for the i-th coordinate of $\tilde{\mathbf{y}}(\mu)$ we have $\tilde{y}_i(\mu) = \frac{\tilde{v}_i}{\mu - \lambda_i}$ (Line 9). Since we are only interested in $\mathbf{y}(\mu) \in \partial\mathcal{B}(r_\perp)$ and $\mathbf{Q}$ is an orthogonal matrix, we must have (Line 7): $\|\mathbf{y}(\mu)\|^2 = \|\tilde{\mathbf{y}}(\mu)\|^2 = \sum_{i=1}^d \tilde{y}_i(\mu)^2 = r_\perp^2$, and hence $\sum_i \frac{\tilde{v}_i^2}{(\mu - \lambda_i)^2} = r_\perp^2$. After multiplying the equation by $\prod_i (\mu - \lambda_i)^2$, we get an equation of the form $p(\mu) = 0$, where p is a polynomial of degree $2d$. We find roots $\mu_1, \ldots, \mu_{2d}$ of the polynomial in time $O(d^2 \log d \cdot \log\log {}^1\!/\!_\varepsilon)$ [29], where ε is the required root precision. For each i, we compute $\mathbf{y}(\mu_i)$ and verify whether it lies in $\mathcal{B}(r_\perp) \cap S_\perp$ and improves the objective by $-\Omega(\delta)$.

Precision. Since the roots of the polynomial are approximate, when μ is close to λ_i, even a small perturbation of μ can strongly affect $y_i(\mu) = \frac{\tilde{v}_i}{\mu - \lambda_i}$. We solve this as follows: since $\|\tilde{\mathbf{y}}(\mu)\| = r_\perp$, for each i we must have $|\tilde{y}_i(\mu)| \le r_\perp$, implying $|\mu - \lambda_i| \ge \frac{|\tilde{v}_i|}{r_\perp}$. Therefore, μ must be sufficiently far from any λ_i, where the lower bound on the distance depends on $r_\perp \le \sqrt[3]{{}^\delta\!/\!_\rho}$ and on $|\tilde{v}_i|$. By adding noise to $\mathbf{v}$ (Line 7 of Algorithm 3), we guarantee that the noise is preserved in $\tilde{\mathbf{v}}$ so that each coordinate is sufficiently separated from 0 w.h.p. This is formalized in Appendix C, Lemma 24.

C.3. Proof of correctness

Algorithm 5: $VRSG(\mathbf{x}, \tilde{\sigma})$: Variance Reduced Stochastic Gradient

parameters: stochastic gradient descent variance σ, error probability ε
output : $\mathbf{g}$ such that $\|\mathbf{g} - \nabla f(\mathbf{x})\| < \tilde{\sigma}$ with probability at least $1 - \varepsilon$

if $\sigma = 0$ **then return** $\nabla f(\mathbf{x})$;
$K \leftarrow \frac{2\sigma^2}{\tilde{\sigma}^2}, \quad M \leftarrow O(\log {}^d\!/\!_\varepsilon)$
for $i = 1, \ldots, M$ **do**
  Sample $\mathbf{g}_{i,1}, \ldots, \mathbf{g}_{i,K}$ – independently sampled stochastic gradients of f at $\mathbf{x}$
  $\mathbf{m}_i \leftarrow \frac{1}{K}\sum_{j=1}^K \mathbf{g}_{i,j}$
return the coordinate-wise median of $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M$

Lemma 20 *Let stochastic gradient oracle satisfy Assumption 2. Then for any $\tilde{\sigma} > 0$, Algorithm 5 returns a vector $\mathbf{g}$ with $\|\mathbf{g} - \nabla f(\mathbf{x})\| < \tilde{\sigma}$ with probability $1 - O(\varepsilon)$ using $O(1 + \frac{\sigma^2}{\tilde{\sigma}^2} \log \frac{d}{\varepsilon})$ stochastic gradient oracle calls.*

Proof Since the stochastic gradients are sampled independently and $\mathbb{E}\|\mathbf{g}_{i,j} - \nabla f(\mathbf{x})\|^2 \le \sigma^2$ by Assumption 2, for each i we have

$$\mathbb{E}\|\mathbf{m}_i - \nabla f(\mathbf{x})\|^2 = \mathbb{E}\|\frac{1}{K}\sum_{j=1}^{k}\mathbf{g}_{i,j} - \nabla f(\mathbf{x})\|^2 \le \frac{\sigma^2}{K} = \frac{\tilde{\sigma}^2}{2}.$$

Applying Fact 11 to $\mathbf{m}_1, \ldots, \mathbf{m}_M$, we have $\|\mathbf{g} - \nabla f(\mathbf{x})\| < \tilde{\sigma}$ with probability

$$1 - d \cdot e^{-\Omega(M)} = 1 - d \cdot e^{-\Omega(\log d/\varepsilon)} = 1 - O(\varepsilon).$$

∎

Lemma 21 (Lemma 19) *Let f satisfy Assumptions 1 and 2. Let $f'(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^\top\mathbf{v} + \frac{1}{2}\mathbf{h}^\top\mathbf{M}\mathbf{h}$, where $\mathbf{v}$ and $\mathbf{M}$ are as in Algorithm 3. For $\delta > 0$, $r = \sqrt[3]{\delta/\rho}$ we have $\|f'(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}+\mathbf{h})\| < \frac{\delta}{2}$ for all $\mathbf{h} \in \mathcal{B}(r)$ w.h.p.*

Proof Recall that $\mathbf{v}$ and $\mathbf{M}$ are noisy estimations of the gradient and the Hessian (below we reiterate how they are computed). By the Taylor expansion of f, we have

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^\top\nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{h}^\top\nabla^2 f(\mathbf{x})\mathbf{h} + \phi(\mathbf{h}),$$

where $\phi(\mathbf{h}) = O(\|\mathbf{h}\|^3)$. To bound $|f'(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}+\mathbf{h})|$, we analyze difference between f and f' for linear, quadratic and higher-order terms separately. Namely, we bound $|\mathbf{h}^\top\nabla f(\mathbf{x}) - \mathbf{h}^\top\mathbf{v}|$, $|\frac{1}{2}\mathbf{h}^\top\nabla^2 f(\mathbf{x})\mathbf{h} - \frac{1}{2}\mathbf{h}^\top\mathbf{M}\mathbf{h}|$, and $|\phi(\mathbf{h})|$ with $\frac{\delta}{6}$ each.

Higher-order terms Since f has a ρ-Lipschitz Hessian, we have $|\phi(\mathbf{h})| \le \frac{\rho r^3}{6} \le \frac{\delta}{6}$ for $r = \sqrt[3]{\delta/\rho}$.

Linear terms Recall that $\mathbf{v} = \mathbf{g} + \zeta$, where $\mathbf{g} = \mathrm{VRSG}(\mathbf{x}, O(\delta/r))$ and $\zeta \sim \mathcal{N}(\mathbf{0}, \xi\frac{\delta}{r\sqrt{d}} \cdot I)$ for some small ξ. By Lemma 20, $\|\mathbf{g} - \nabla f(\mathbf{x})\| < O(\delta/r)$ w.h.p. By choosing the appropriate constants, with probability $1 - O(\varepsilon)$ we have

$$\|\mathbf{v} - \nabla f(\mathbf{x})\| = \|\mathbf{g} - \nabla f(\mathbf{x})\| + \|\zeta\| \le \frac{\delta}{12r} + \xi\frac{\delta}{r}\log\frac{1}{\varepsilon} \le \frac{\delta}{6r},$$

where we selected $\xi \le \frac{1}{12}\log^{-1}\frac{1}{\varepsilon}$. Using this bound, for all $\mathbf{h} \in \mathcal{B}(r)$ we have:

$$|\mathbf{h}^\top\nabla f(\mathbf{x}) - \mathbf{h}^\top\mathbf{v}| \le r\|\mathbf{v} - \nabla f(\mathbf{x})\| < \frac{\delta}{6}$$

Quadratic terms Recall that $\mathbf{M} = \mathbf{H} + \gamma I$, where $\gamma \sim U[-\xi, \xi]$ and $\mathbf{H}$ is as in Algorithm 3: for $\{\mathbf{e}_i\}$ – the standard basis in $\mathbb{R}^d$, $\theta = \Theta(\frac{\delta}{\rho r^2})$ and $\tilde{\sigma} = \Theta(\frac{\delta^2}{\rho r^2})$, the i-th column of $\mathbf{H}$ is $\frac{\mathbf{g}_i - \mathbf{g}^{(0)}}{\theta}$ with $\mathbf{g}_i = \mathrm{VRSG}(\mathbf{x} + \theta\mathbf{e}_i, \tilde{\sigma})$ and $\mathbf{g}^{(0)} = \mathrm{VRSG}(\mathbf{x} + \theta\mathbf{e}_i, \tilde{\sigma})$. For all $\mathbf{h} \in \mathcal{B}(r)$ we have:

$$|\frac{1}{2}\mathbf{h}^\top\mathbf{M}\mathbf{h} - \frac{1}{2}\mathbf{h}^\top\nabla^2 f(\mathbf{x})\mathbf{h}| \le |\frac{1}{2}\mathbf{h}^\top(\mathbf{H} - \nabla^2 f(\mathbf{x}))\mathbf{h}| + \frac{|\gamma|}{2}\|\mathbf{h}\|^2.$$

The second term is bounded by selecting a sufficiently small γ (i.e. a sufficiently small ξ), and hence it remains to bound the second term. Let $\mathbf{w}_i$ be the difference between $\frac{\nabla f(\mathbf{x}+\theta\mathbf{e}_i)-\nabla f(\mathbf{x})}{\theta}$ and its stochastic estimate:

$$\mathbf{w}_i = \frac{(\mathbf{g}_i - \mathbf{g}^{(0)}) - (\nabla f(\mathbf{x}+\theta\mathbf{e}_i) - \nabla f(\mathbf{x}))}{\theta}.$$

Since $\mathbf{g}_i$ and $\mathbf{g}^{(0)}$ are within distance $\tilde{\sigma}$ of the true gradients, we have $\|\mathbf{w}_i\| = O\left(\frac{\tilde{\sigma}}{\theta}\right)$. Since $\nabla^2 f(\mathbf{x})\mathbf{e}_i$ is the i-th column of $\nabla^2 f(\mathbf{x})$, we have:

$$\begin{aligned}
&\frac{\mathbf{g}_i - \mathbf{g}^{(0)}}{\theta} - \nabla^2 f(\mathbf{x})\mathbf{e}_i \\
&= \mathbf{w}_i + \frac{\nabla f(\mathbf{x}+\theta\mathbf{e}_i) - \nabla f(\mathbf{x})}{\theta} - \nabla^2 f(\mathbf{x})\mathbf{e}_i \\
&= \mathbf{w}_i + \int_{\tau=0}^{1} (\nabla^2 f(\mathbf{x}+\tau\theta\mathbf{e}_i) - \nabla^2 f(\mathbf{x})) d\tau \cdot \mathbf{e}_i \\
&= \mathbf{w}_i + \mathbf{u}_i,
\end{aligned}$$

where $\|\mathbf{u}_i\| = O(\rho\theta)$ since $\nabla^2 f$ is ρ-Lipschitz. Therefore, for any $\mathbf{h} \in \mathcal{B}(r)$ we have

$$(\mathbf{H} - \nabla^2 f(\mathbf{x}))\mathbf{h} = \sum_{i=1}^{d} h_i(\mathbf{w}_i + \mathbf{u}_i),$$

and hence:

$$\begin{aligned}
|\mathbf{h}^\top(\mathbf{H} - \nabla^2 f(\mathbf{x}))\mathbf{h}| &= |\sum_{i=1}^{d} h_i(\mathbf{w}_i + \mathbf{u}_i)^\top \mathbf{h}| \\
&\le \sum_{i=1}^{d} |h_i| \cdot \|\mathbf{w}_i + \mathbf{u}_i\| \cdot \|\mathbf{h}\| \\
&= O\left(\|\mathbf{h}\|_1 (\frac{\tilde{\sigma}}{\theta} + \rho\theta)\|\mathbf{h}\|\right) \\
&= O\left(\sqrt{d} r^2 (\frac{\tilde{\sigma}}{\theta} + \rho\theta)\right),
\end{aligned}$$

where the third line follows from $\|\mathbf{w}_i\| = O(\frac{\tilde{\sigma}}{\theta})$ and $\|\mathbf{u}_i\| = O(\rho\theta)$). Hence, it suffices to choose θ so that:

$$\sqrt{d} r^2 \rho\theta = \Theta(\delta) \iff \theta = \Theta(\frac{\delta}{\sqrt{d}\rho r^2}) = \Theta(\frac{r}{\sqrt{d}}),$$

where we used $r = \sqrt[3]{\delta/\rho}$. Similarly, we choose $\tilde{\sigma}$ so that:

$$\sqrt{d} r^2 \frac{\tilde{\sigma}}{\theta} = \Theta(\delta) \iff \tilde{\sigma} = \Theta(\frac{\delta\theta}{\sqrt{d} r^2}) = \Theta(\frac{\rho r^2}{d}).$$

■

Lemma 22 *Let $\delta, r > 0$. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x} + \mathbf{x}^\top \mathbf{v} + C$, with the following conditions satisfied:*

- $\lambda_{\max}(\mathbf{M}) \leq L$
- *(Large linear term)* $\|\mathbf{v}\| \geq 4(Ldr + \frac{(C+\delta)d}{r})$
- *For* $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x} \in S \cap \mathcal{B}(r)} f(\mathbf{x})$ *we have* $f(\mathbf{x}^\star) < -\delta$

Then either $(\mathbf{x}^\star)^\top \mathbf{v} < -\frac{r}{8d}\|\mathbf{v}\|$ *or* $\mathbf{x}^\star \in \partial S$.

Remark 23 *As we show in the proof of Lemma 24, if* $\|\mathbf{v}\| \geq Ldr + \frac{(C+\delta)d}{r}$ *and* $\mathbf{x}^\star \notin \partial S$*, the linear term alone is sufficient to sufficiently improve the function, and the quadratic term is negligible.*

Proof For contradiction, assume that $(\mathbf{x}^\star)^\top \mathbf{v} \geq -\frac{r}{4d}\|\mathbf{v}\|$ and $\mathbf{x}^\star \in \operatorname{Int} S$. We show that we can perturb $\mathbf{x}^\star$ so that the point still lies in $\mathcal{B}(r) \cap S$ and the objective decreases. We consider two cases: when $\mathbf{x}^\star \in \operatorname{Int} \mathcal{B}(r)$, we can simply move in the direction of $-\mathbf{v}$. When $\mathbf{x}^\star \in \partial \mathcal{B}(r)$, we rely on the fact that $(\mathbf{x}^\star)^\top \mathbf{v} < -\frac{r}{8d}\|\mathbf{v}\|$, which guarantees that small change in direction $-\mathbf{v}$ requires a small change in the orthogonal direction.

Case 1: $\mathbf{x}^\star \in \operatorname{Int} \mathcal{B}(r)$**.** Let $\mathbf{x}^\star = \alpha\mathbf{v} + \mathbf{w}$, where $\mathbf{w} \perp \mathbf{v}$. Since $\mathbf{x}^\star \in \operatorname{Int}(\mathcal{B}(r) \cap S)$, for a sufficiently small ε we have $\mathbf{x}^\star - \varepsilon\mathbf{v} \in \mathcal{B}(r) \cap S$, and then:

$$
\begin{aligned}
&f(\mathbf{x}^\star - \varepsilon\mathbf{v}) \\
&= \frac{1}{2}(\mathbf{x}^\star - \varepsilon\mathbf{v})^\top \mathbf{M}(\mathbf{x}^\star - \varepsilon\mathbf{v}) + (\mathbf{x}^\star - \varepsilon\mathbf{v})^\top \mathbf{v} + C \\
&= f(\mathbf{x}^\star) - \varepsilon(\mathbf{x}^\star)^\top \mathbf{M}\mathbf{v} + \varepsilon^2 \frac{1}{2}\mathbf{v}^\top \mathbf{M}\mathbf{v} - \varepsilon\|\mathbf{v}\|^2 \\
&\leq f(\mathbf{x}^\star) + \varepsilon(rL\|\mathbf{v}\| - \|\mathbf{v}\|^2) + O(\varepsilon^2) \\
&< f(\mathbf{x}^\star),
\end{aligned}
$$

contradicting the fact that $\mathbf{x}^\star$ is the minimizer.

Case 2: $\mathbf{x}^\star \in \partial\mathcal{B}(r)$**.** Let $\mathbf{x}^\star = -\alpha\mathbf{v} + \mathbf{w}$, where $\mathbf{w} \perp \mathbf{v}$. Since $(\mathbf{x}^\star)^\top \mathbf{v} \geq -\frac{r}{8d}\|\mathbf{v}\|$, we have $\alpha \leq \frac{r}{8d\|\mathbf{v}\|}$, and hence $\|\mathbf{w}\| \geq \frac{r}{2}$ and $\alpha\|\mathbf{v}\| \leq \frac{\|w\|}{4d}$. We'll construct $\mathbf{x}'$ such that $\mathbf{x}' \in \mathcal{B}(r)$, $\mathbf{x}'$ is arbitrarily close to $\mathbf{x}^\star$, and, since $\mathbf{x}^\star \in \operatorname{Int} S$, we also have $\mathbf{x}' \in \operatorname{Int} S$. Let $\mathbf{x}' = -(\alpha+\varepsilon)\mathbf{v} + (1-\theta)\mathbf{w}$ for $\varepsilon > 0$. To have $\mathbf{x}' \in \mathcal{B}(r)$, it suffices to guarantee that $\|\mathbf{x}'\| \leq \|\mathbf{x}^\star\|$, meaning:

$$
\begin{aligned}
&(\alpha + \varepsilon)^2\|\mathbf{v}\|^2 + (1-\theta)^2\|\mathbf{w}\|^2 \leq \alpha^2\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 \\
&\iff (2\alpha\varepsilon + \alpha\varepsilon^2)\|\mathbf{v}\|^2 \leq \theta(2-\theta)\|\mathbf{w}\|^2,
\end{aligned}
$$

and hence for $\mathbf{x}^\star \in \partial\mathcal{B}(r)$ it suffices to select $\theta = \frac{2\alpha\varepsilon\|\mathbf{v}\|^2}{\|\mathbf{w}\|^2}$. Since $\theta = O(\varepsilon)$, $\|\mathbf{x}' - \mathbf{x}^\star\| = O(\varepsilon)$, and hence we can select $\mathbf{x}'$ arbitrarily close to $\mathbf{x}^\star$, which guarantees $\mathbf{x}' \in S$. It remains to show that

$f(\mathbf{x}') < f(\mathbf{x}^\star)$. Using $\theta = \frac{2\alpha\varepsilon\|\mathbf{v}\|^2}{\|\mathbf{w}\|^2} \le \frac{\varepsilon}{2d}\frac{\|\mathbf{v}\|}{\|\mathbf{w}\|}$:

$$\begin{aligned}
f(\mathbf{x}') &= f(-(\alpha+\varepsilon)\mathbf{v} + (1-\theta)\mathbf{w}) \\
&= \frac{1}{2}(-(\alpha+\varepsilon)\mathbf{v} + (1-\theta)\mathbf{w})^\top \mathbf{M}(-(\alpha+\varepsilon)\mathbf{v} \\
&\quad + (1-\theta)\mathbf{w}) + ((\alpha+\varepsilon)\mathbf{v} + (1-\theta)\mathbf{w})^\top \mathbf{v} + C \\
&= f(\mathbf{x}^\star) + \frac{1}{2}(-\varepsilon\mathbf{v} - \theta\mathbf{w})^\top \mathbf{M}(-\varepsilon\mathbf{v} - \theta\mathbf{w}) \\
&\quad + (-\varepsilon\mathbf{v} - \theta\mathbf{w})^\top \mathbf{M}\mathbf{x}^\star + (-\varepsilon\mathbf{v} - \theta\mathbf{w})^\top \mathbf{v} \\
&\le f(\mathbf{x}^\star) + (\varepsilon + \frac{\varepsilon}{2d})\|\mathbf{v}\|Lr + O(\varepsilon^2) - \varepsilon\|\mathbf{v}\|^2(\varepsilon - \frac{\varepsilon}{2d}) \\
&< f(\mathbf{x}^\star).
\end{aligned}$$

■

Lemma 24 *For $\delta, r, \theta > 0$, let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{y} \ge \mathbf{b}\}$ and let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}\mathbf{x} + \mathbf{x}^\top\mathbf{v} + C$ satisfy the following conditions:*

1. *$\lambda_{|max|}(\mathbf{M}) \le L$,*
2. *$(\mathbf{M} - \mu I)\mathbf{y} = -\mathbf{v}$ has no solutions for all μ such that $\operatorname{rank}(\mathbf{\Lambda} - \mu I) < d$,*
3. *$|\tilde{v}_i| \ge \theta$ for all i, where $\tilde{\mathbf{v}} = \mathbf{Q}\mathbf{v}$ and $\mathbf{Q}$ is an orthogonal matrix defined in Line 5 of Algorithm 4.*

Let $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{y} \in S \cap \mathcal{B}(r)} f(\mathbf{x})$ and assume that $f(\mathbf{x}^\star) < -\delta$ and $\mathbf{x}^\star \in \operatorname{Int} S$. Then Algorithm 4 finds $\mathbf{x} \in S \cap \mathcal{B}(r)$ such that $f(\mathbf{x}) < -(1-\varepsilon)\delta$ in time $\tilde{O}(d^3)$.

Remark 25 *In the statement above, we select θ sufficiently small so that adding random noise to $\nabla f(\mathbf{0})$ guarantees that condition 3 is satisfied w.h.p. Condition 2 is not required, but it's easy to satisfy and simplifies the analysis.*

Proof As outlined in Section 2.2, we analyze the cases from Algorithm 3 separately. We consider the situation when $\|\mathbf{v}\|$ is large as a separate case, since in this case the function is sensitive to the change of its argument. Otherwise, for $\mathbf{x}^\star$, there are only two options: either $\mathbf{x}^\star \in \operatorname{Int} \mathcal{B}(r)$ or $\mathbf{x}^\star \in \partial\mathcal{B}(r)$.

Case 1: Large $\|\mathbf{v}\|$. If $\|\mathbf{v}\| \ge 4(Ldr + \frac{(C+\delta)d}{r})$, then by Lemma 22 either $(\mathbf{x}^\star)^\top\mathbf{v} < -\frac{r}{4d}\|\mathbf{v}\|$ or $\mathbf{x}^\star \in \partial S$. Since we assume that $\mathbf{x}^\star \in \operatorname{Int} S$, it leaves us with the case $(\mathbf{x}^\star)^\top\mathbf{v} < -\frac{r}{4d}\|\mathbf{v}\|$. Then Algorithm 3 finds $\mathbf{x}$ with $\mathbf{x}^\top\mathbf{v} < -\frac{r}{d}\|\mathbf{v}\|$ and hence

$$\begin{aligned}
f(\mathbf{x}) &\le \frac{1}{2}Lr^2 + C - \frac{r}{4d}\|\mathbf{v}\|^2 \\
&\le \frac{1}{2}Lr^2 + C - \frac{r}{4d}4(Ldr + (C+\delta)\frac{d}{r}) \\
&\le -\delta.
\end{aligned}$$

Case 2: $\mathbf{x}^\star \in \operatorname{Int} S \cap \operatorname{Int} \mathcal{B}(r)$. In this case, $\mathbf{x}^\star$ is an unconstrained critical point, and so we must have $\nabla f(\mathbf{x}) = \mathbf{M}\mathbf{x} + \mathbf{v} = \mathbf{0}$. If $\mathbf{M}$ is degenerate, by condition 2, there are no solutions. Otherwise, there exists a unique $\mathbf{x}$ such that $\mathbf{M}\mathbf{x} = -\mathbf{v}$. We then check that $\mathbf{x} \in S \cap \mathcal{B}(r)$ and $f(\mathbf{x}) < f(\mathbf{0}) - \frac{\delta}{2}$.

Case 3: $\mathbf{x}^\star \in \operatorname{Int} S \cap \partial\mathcal{B}(r)$. By the method of Lagrangian multipliers, we must have $\mathbf{M}\mathbf{x}+\mathbf{v} = \mu\mathbf{x}$ for some λ. By condition 2, if $\operatorname{rank}(\mathbf{M} - \mu I) < d$, then $(\mathbf{M} - \mu I)\mathbf{x} = -\mathbf{v}$ has no solutions. Hence, it suffices to consider cases when μ is not an eigenvalue of $\mathbf{M}$.

When μ is not an eigenvalue of $\mathbf{M}$, $(\mathbf{M} - \mu I)\mathbf{x} = -\mathbf{v}$ has a unique solution. Based on [30], we find orthogonal $\mathbf{Q}$ and diagonal $\mathbf{\Lambda}$ such that $\|\mathbf{M} - \mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}\| < \varepsilon$, which can be done in time $O(d^3 \log {}^1/\varepsilon)$. By setting $\varepsilon = O({}^\delta/r^2)$, we guarantee that the function changes by at most $O(\delta)$ in $\mathcal{B}(r)$. Then $(\mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q} - \mu I)\mathbf{x} = -\mathbf{v}$ is equivalent to

$$(\mathbf{\Lambda} - \mu I)\mathbf{Q}\mathbf{x} = -\mathbf{Q}\mathbf{v}.$$

Let $\tilde{\mathbf{v}} = \mathbf{Q}\mathbf{v}$ and $\tilde{\mathbf{x}} = \mathbf{Q}\mathbf{x}$. Then the solutions for $\tilde{\mathbf{x}}$ have coordinates $\tilde{x}_i = \frac{-\tilde{v}_i}{\mu-\lambda_i}$ for all i. Since we are only interested in solutions on $\partial\mathcal{B}(r)$ and $\mathbf{Q}$ is an orthogonal matrix, we must have

$$\|\mathbf{x}\|^2 = \|\tilde{\mathbf{x}}\|^2 = \sum_{i=1}^{d} \frac{\tilde{v}_i^2}{(\mu - \lambda_i)^2} = r^2.$$

By multiplying both sides by $\prod_i(\mu - \lambda_i)^2$, we get an equation of form $p(\lambda) = 0$, where p is the polynomial with degree $2d$ and the highest degree coefficient being r^2. We can find all the roots of p with precision ν (to be specified later) in time $O(d^2 \log d \cdot \log\log \frac{1}{\nu})$ [29]. Since we are only interested in real roots, we take the real parts of the approximate roots (if the exact root is real, taking the real part only improves the accuracy).

Next, we estimate the required precision ν. Since we have $\|\tilde{\mathbf{x}}\| = r$, it requires, in particular, that $|x_i| \le r$ for all i, which means that λ_i must satisfy

$$\frac{|\tilde{v}_i|}{|\mu - \lambda_i|} \le r \iff |\mu - \lambda_i| \ge \frac{|\tilde{v}_i|}{r}.$$

Since $|v_i| \ge \theta$ by our assumption, we have $|\mu - \lambda_i| \ge \frac{\theta}{r}$.

Let μ be the exact root of the polynomial and $\mu' \in [\mu-\nu, \mu+\nu]$ be the corresponding approximate root found the algorithm. Let $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ be the points in $\mathbb{R}^d$ corresponding to μ and μ'. Then we have:

$$\begin{aligned}
|\tilde{x}_i - \tilde{x}_i'| &= |\frac{\tilde{v}_i}{\mu - \lambda_i} - \frac{\tilde{v}_i}{\mu' - \lambda_i}| \\
&= |\frac{\tilde{v}_i}{\mu - \lambda_i} \cdot \frac{\mu - \mu'}{\mu' - \lambda_i}| \\
&\le \frac{r\nu}{|\mu' - \lambda_i|} \\
&\le \frac{r\nu}{\frac{\theta}{r} - \nu} \\
&\le \frac{2r^2\nu}{\theta},
\end{aligned}$$

where we assumed $\nu < \frac{\theta}{2r}$. If $\mathbf{x} = \mathbf{Q}^\top\tilde{\mathbf{x}}$ and $\mathbf{x}' = \mathbf{Q}^\top\tilde{\mathbf{x}}$, then $\|\mathbf{x} - \mathbf{x}'\| = \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}'\| \le \frac{2\sqrt{d}r^2\nu}{\theta}$.

Let $\mathbf{x}''$ be the projection of $\mathbf{x}'$ on $\mathcal{B}(r) \cap S$. Then $\|\mathbf{x}'' - \mathbf{x}\| \le 2\|\mathbf{x}' - \mathbf{x}\|$, and hence:

$$\begin{aligned}
|f(\mathbf{x}) - f(\mathbf{x}'')| &= \left|\frac{1}{2}(\mathbf{x}^\top \mathbf{M}\mathbf{x} - \mathbf{x}''^\top \mathbf{M}\mathbf{x}'') + (\mathbf{x} - \mathbf{x}'')^\top \mathbf{v}\right| \\
&= \left|\frac{1}{2}(\mathbf{x} - \mathbf{x}'')^\top \mathbf{M}(\mathbf{x} + \mathbf{x}'') + (\mathbf{x} - \mathbf{x}'')^\top \mathbf{v}\right| \\
&\le \|\mathbf{x} - \mathbf{x}''\|(Lr + \|\mathbf{v}\|) \\
&\le \nu \cdot \frac{4\sqrt{d}r^2}{\theta}(L(d+1)r + \frac{(C+\delta)d}{r}) \\
&\le \frac{\delta}{10},
\end{aligned}$$

by selecting $\nu < \frac{\delta\theta}{40\sqrt{d}(L(d+2)r^3 + (C+\delta)dr)}$. ■

Theorem 26 (Theorem 8) *Let $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ be a set defined by an intersection of k linear inequality constraints. Let f satisfy Assumptions 1 and 2 and let $\min_{\mathbf{x}\in S} f(\mathbf{x}) = f^\star$. Then there exists an algorithm which for $\delta > 0$ finds a δ-SOSP in $\tilde{O}(\frac{f(\mathbf{x}_0) - f^\star}{\delta} d^3(2^k + \frac{\sigma^2}{\delta^{4/3}}))$ time using $\tilde{O}(\frac{f(\mathbf{x}_0) - f^\star}{\delta}(d + \frac{d^3\sigma^2}{\delta^{4/3}}))$ stochastic gradient oracle calls w.h.p.*

Proof It suffices to show that if $\mathbf{x}$ is not a δ-SOSP, then Algorithm 3 decreases function value by $\Omega(\delta)$, and hence there can't be more that $O(\frac{f(\mathbf{x}_0) - f^\star}{\delta})$ iterations. W.l.o.g. we can assume that the current iterate is $\mathbf{0}$ and $f(\mathbf{0}) = 0$. As shown in Lemma 19, we can replace f with its quadratic approximation $\frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x} + \mathbf{x}^\top \mathbf{v}$ while increasing the function value by at most $\frac{\delta}{2}$.

As shown in Section C.1, we can consider function $g(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top \mathbf{M}_\perp \mathbf{y} + \mathbf{y}^\top \mathbf{v}_\perp + C$ in $S_\perp \cap \mathcal{B}(r_\perp)$. By Lemma 24, we can find $\mathbf{y}$ with $g(\mathbf{y}) < -\frac{1}{2}\delta$ in $\tilde{O}(d^3)$ time. It remains to show that the conditions from Lemma 24 are satisfied. Recall that for $\mathbf{O} \in \mathbb{R}^{d \times \dim \mathcal{A}}$ – an orthonormal basis of $\mathcal{A}$, we define the following:

- $S_\perp \leftarrow \{\mathbf{y} \mid \mathbf{A}(\mathbf{p} + \mathbf{O}\mathbf{y}) \le \mathbf{b}\}$
- $\mathbf{M}_\perp \leftarrow \mathbf{O}^\top \mathbf{M}\mathbf{O}$
- $\mathbf{v}_\perp \leftarrow \mathbf{O}^\top(\mathbf{v} + \mathbf{M}(\mathbf{p} - \mathbf{x}))$

As required in Lemma 24, $S_\perp$ is the set defined by linear inequality constraints. We next check other requirements in Lemma 24.

1) $\lambda_{|max|}(\mathbf{M}_\perp) \le L$: Since $\mathbf{O}$ defines an orthonormal basis, $\lambda_{|max|}(\mathbf{M}_\perp) = \lambda_{|max|}(\mathbf{O}^\top \mathbf{M}\mathbf{O}) \le \lambda_{|max|}(\mathbf{M}) \le L$.

2) $(\mathbf{M}_\perp - \mu I)\mathbf{y} = -\mathbf{v}_\perp$ has no solutions for all μ such that $\operatorname{rank}(\mathbf{\Lambda} - \mu I) < d$

$$\mathbf{v}_\perp = \mathbf{O}^\top(\mathbf{v} + \mathbf{M}(\mathbf{p} - \mathbf{x})) = \mathbf{O}^\top\zeta + \mathbf{O}^\top(\mathbf{g} + \mathbf{M}(\mathbf{p} - \mathbf{x})),$$

where $\zeta \sim \mathcal{N}(\mathbf{0}, \xi\frac{\delta}{r\sqrt{d}} \cdot I)$. Since $\mathbf{O}$ is an orthonormal basis, we have $\mathbf{O}^\top\zeta \sim \mathcal{N}(\mathbf{0}, \xi\frac{\delta}{r\sqrt{d}} \cdot I)$. This Gaussian noise guarantees that the $\mathbf{v}_\perp$ has the probability 0 to lie in any fixed subspace, which guarantees that there are no solutions for degenerate systems of the form $(\mathbf{M}_\perp - \mu I)\mathbf{y} = -\mathbf{v}_\perp$.

3) $|\tilde{v}_i| \geq \theta$ **for all** i Recall that $\tilde{\mathbf{v}} = \mathbf{Q}\mathbf{v}_\perp$ as defined in Line 6 of Algorithm 4, where $\mathbf{Q}$ is an orthogonal matrix. For a fixed i and θ, we consider $\Pr_\zeta[|\tilde{v}_i| < \theta]$. Due to rotation invariance of Gaussian noise, its projection on any linear subspace has distribution $\mathcal{N}(\mathbf{0}, \xi\frac{\delta}{r\sqrt{d}}I)$. By selecting $\theta = \varepsilon\frac{\xi\delta}{d^2 r}$, we have that the for all i, $|\tilde{v}_i| > \theta$ with probability at least $1-\varepsilon$.

By selecting ε such that taking the union bound over all $O(\frac{f(\mathbf{x}_0)-f^\star}{\delta})$ iterations, d coordinates and 2^k sets of constraints gives bounded error probability, we finish the proof. ■