JMLR: Workshop and Conference Proceedings vol 40:1–46, 2015

Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition

Rong Ge RONGGE@MICROSOFT.COM
Microsoft Research New England

Furong Huang FURONGH@UCI.EDU
University of California Irvine

Chi Jin CHIJIN@CS.BERKELEY.EDU
University of California Berkeley

Yang Yuan YANGYUAN@CS.CORNELL.EDU
Cornell University

Abstract

We analyze stochastic gradient descent for optimizing non-convex functions. In many cases for non-convex functions the goal is to find a reasonable local minimum, and the main concern is that gradient updates are trapped in *saddle points*. In this paper we identify *strict saddle* property for non-convex problem that allows for efficient optimization. Using this property we show that from an *arbitrary* starting point, stochastic gradient descent converges to a local minimum in a polynomial number of iterations. To the best of our knowledge this is the first work that gives *global* convergence guarantees for stochastic gradient descent on non-convex functions with exponentially many local minima and saddle points.

Our analysis can be applied to orthogonal tensor decomposition, which is widely used in learning a rich class of latent variable models. We propose a new optimization formulation for the tensor decomposition problem that has strict saddle property. As a result we get the first online algorithm for orthogonal tensor decomposition with global convergence guarantee.

Keywords: stochastic gradient, non-convex optimization, saddle points, tensor decomposition

1. Introduction

Stochastic gradient descent is one of the basic algorithms in optimization. It is often used to solve the following stochastic optimization problem

$$w = \arg\min_{w \in \mathbb{R}^d} f(w), \text{ where } f(w) = \mathbb{E}_{x\sim\mathcal{D}}[\phi(w, x)] \tag{1}$$

Here x is a data point that comes from some unknown distribution $\mathcal{D}$, and ϕ is a *loss function* that is defined for a pair (x, w) of sample and parameters. We hope to minimize the expected loss $\mathbb{E}[\phi(w, x)]$.

When the function $f(w)$ is convex, convergence of stochastic gradient descent is well-understood (Shalev-Shwartz et al., 2009; Rakhlin et al., 2012). However, stochastic gradient descent is not only limited to convex functions. Especially, in the context of neural networks, stochastic gradient descent is known as the "backpropagation" algorithm (Rumelhart et al., 1988), and has been the main algorithm that underlies the success of deep learning (Bengio, 2009). However, the guarantees in the convex setting does not transfer to the non-convex settings.

Optimizing a non-convex function is NP-hard in general. The difficulty comes from two aspects. First, the function may have many local minima, and it might be hard to find the best one (global minimum) among them. Second, even finding a local minimum might be hard as there can be many saddle points which have 0-gradient but are not local minima[1]. In the most general case, there is no known algorithm that guarantees to find a local minimum in polynomial number of steps. The discrete analog (finding local minimum in domains like $\{0,1\}^n$) has been studied in complexity theory and is PLS-complete (Johnson et al., 1988).

In many cases, especially in those related to deep neural networks (Dauphin et al., 2014) (Choromanska et al., 2014), the main bottleneck in optimization is not due to local minima, but the existence of many saddle points. Gradient based algorithms are in particular susceptible to saddle point problems as they only rely on the gradient information. The saddle point problem is alleviated for second-order methods that also rely on the Hessian information (Dauphin et al., 2014).

However, using Hessian information usually increases the memory requirement and computation time per iteration. As a result many applications still use stochastic gradient and empirically get reasonable results. In this paper we investigate why stochastic gradient methods can be effective even in presence of saddle point, in particular we answer the following question:

Question: Given a non-convex function f with many saddle points, what properties of f will guarantee stochastic gradient descent to converge to a local minimum efficiently?

We identify a property of non-convex functions which we call *strict saddle*. Intuitively, it guarantees local progress if we have access to the Hessian information. Surprisingly we show with only first order (gradient) information, stochastic gradient can escape from the saddle points efficiently. We give a framework for analyzing stochastic gradient in both unconstrained and equality-constrained case using this property.

We apply our framework to *orthogonal tensor decomposition*, which is a core problem in learning many latent variable models (see discussion in Section 2.2). The tensor decomposition problem is inherently susceptible to the saddle point issues, as the problem asks to find d different components and any permutation of the true components yields a valid solution. Such symmetry creates exponentially many local minima and saddle points in the optimization problem. Using our new analysis of stochastic gradient, we give the first online algorithm for orthogonal tensor decomposition with global convergence guarantee. This is a key step towards making tensor decomposition algorithms more scalable.

1.1. Summary of Results

Strict saddle functions Given a function $f(w)$ that is twice differentiable, we call w a stationary point if $\nabla f(w) = 0$. A stationary point can either be a local minimum, a local maximum or a saddle point. We identify an interesting class of non-convex functions which we call strict saddle. For these functions the Hessian of every saddle point has a negative eigenvalue. In particular, this means that local second-order algorithms which are similar to the ones in (Dauphin et al., 2014) can always make some progress.

It may seem counter-intuitive why stochastic gradient can work in these cases: in particular if we run the basic gradient descent starting from a stationary point then it will not move. However, we show that the saddle points are not stable and that the randomness in stochastic gradient helps the algorithm to escape from the saddle points.

Theorem 1 (informal) *Suppose $f(w)$ is strict saddle (see Definition 5), Noisy Gradient Descent (Algorithm 1) outputs a point that is close to a local minimum in polynomial number of steps.*

Online tensor decomposition Requiring all saddle points to have a negative eigenvalue may seem strong, but it already allows non-trivial applications to natural non-convex optimization problems. As an example,

1. See Section 3 for definition of saddle points.

we consider the orthogonal tensor decomposition problem. This problem is the key step in spectral learning for many latent variable models (see more discussions in Section 2.2).

We design a new objective function for tensor decomposition that is strict saddle.

Theorem 2 *Given random variables X such that $T = \mathbb{E}[g(X)] \in \mathbb{R}^{d^4}$ is an orthogonal 4-th order tensor (see Section 2.2), there is an objective function $f(w) = \mathbb{E}[\phi(w, X)]$ $w \in \mathbb{R}^{d \times d}$ such that every local minimum of $f(w)$ corresponds to a valid decomposition of T. Further, function f is strict saddle.*

Combining this new objective with our framework for optimizing strict saddlefunctions, we get the first online algorithm for orthogonal tensor decomposition with global convergence guarantee.

1.2. Related Works

Relaxed notions of convexity In optimization theory and economics, there are extensive works on understanding functions that behave similarly to convex functions (and in particular can be optimized efficiently). Such notions involve pseudo-convexity (Mangasarian, 1965), quasi-convexity (Kiwiel, 2001), invexity(Hanson, 1999) and their variants. More recently there are also works that consider classes that admit more efficient optimization procedures like RSC (restricted strong convexity) (Agarwal et al., 2010). Although these classes involve functions that are non-convex, the function (or at least the function restricted to the region of analysis) still has a unique stationary point that is the desired local/global minimum. Therefore these works cannot be used to prove global convergence for problems like tensor decomposition, where by symmetry of the problem there are exponentially many local minima and saddle points.

Second-order algorithms The most popular second-order method is the Newton's method. Although Newton's method converges fast near a local minimum, its global convergence properties are less understood in the more general case. For non-convex functions, (Frieze et al., 1996) gave a concrete example where second-order method converges to the desired local minimum in polynomial number of steps (interestingly the function of interest is trying to find one component in a 4-th order orthogonal tensor, which is a simpler case of our application). As Newton's method often converges also to saddle points, to avoid this behavior, different trusted-region algorithms are applied (Dauphin et al., 2014).

Stochastic gradient and symmetry The tensor decomposition problem we consider in this paper has the following symmetry: the solution is a set of d vectors $v_1, ..., v_d$. If $(v_1, v_2, ..., v_d)$ is a solution, then for any permutation π and any sign flips $\kappa \in \{\pm 1\}^d$, $(.., \kappa_i v_{\pi(i)}, ...)$ is also a valid solution. In general, symmetry is known to generate saddle points, and variants of gradient descent often perform reasonably in these cases (see (Saad and Solla, 1995), (Rattray et al., 1998), (Inoue et al., 2003)). The settings in these work are different from ours, and none of them give bounds on number of steps required for convergence.

There are many other problems that have the same symmetric structure as the tensor decomposition problem, including the sparse coding problem (Olshausen and Field, 1997) and many deep learning applications (Bengio, 2009). In these problems the goal is to learn multiple "features" where the solution is invariant under permutation. Note that there are many recent papers on iterative/gradient based algorithms for problems related to matrix factorization (Jain et al., 2013; Saxe et al., 2013). These problems often have very different symmetry, as if $Y = AX$ then for any invertible matrix R we know $Y = (AR)(R^{-1}X)$. In this case all the equivalent solutions are in a connected low dimensional manifold and there need not be saddle points between them.

2. Preliminaries

Notation Throughout the paper we use $[d]$ to denote set $\{1, 2, ..., d\}$. We use $\|\cdot\|$ to denote the ℓ_2 norm of vectors and spectral norm of matrices. For a matrix we use λ_{min} to denote its smallest eigenvalue. For a function $f : \mathbb{R}^d \to \mathbb{R}$, ∇f and $\nabla^2 f$ denote its gradient vector and Hessian matrix.

2.1. Stochastic Gradient Descent

The stochastic gradient aims to solve the stochastic optimization problem (1), which we restate here:

$$w = \arg\min_{w \in \mathbb{R}^d} f(w), \text{ where } f(w) = \mathbb{E}_{x \sim \mathcal{D}}[\phi(w, x)].$$

Recall $\phi(w, x)$ denotes the loss function evaluated for sample x at point w. The algorithm follows a stochastic gradient

$$w_{t+1} = w_t - \eta \nabla_{w_t} \phi(w_t, x_t), \tag{2}$$

where x_t is a random sample drawn from distribution $\mathcal{D}$ and η is the *learning rate*.

In the more general setting, stochastic gradient descent can be viewed as optimizing an arbitrary function $f(w)$ given a stochastic gradient oracle.

Definition 3 *For a function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a function $SG(w)$ that maps a variable to a random vector in $\mathbb{R}^d$ is a stochastic gradient oracle if $\mathbb{E}[SG(w)] = \nabla f(w)$ and $\|SG(w) - \nabla f(w)\| \leq Q$.*

In this case the update step of the algorithm becomes $w_{t+1} = w_t - \eta SG(w_t)$.

Smoothness and Strong Convexity Traditional analysis for stochastic gradient often assumes the function is smooth and strongly convex. A function is β-smooth if for any two points w_1, w_2,

$$\|\nabla f(w_1) - \nabla f(w_2)\| \leq \beta \|w_1 - w_2\|. \tag{3}$$

When f is twice differentiable this is equivalent to assuming that the spectral norm of the Hessian matrix is bounded by β. We say a function is α-strongly convex if the Hessian at any point has smallest eigenvalue at least α ($\lambda_{min}(\nabla^2 f(w)) \geq \alpha$).

Using these two properties, previous work (Rakhlin et al., 2012) shows that stochastic gradient converges at a rate of $1/t$. In this paper we consider non-convex functions, which can still be β-smooth but cannot be strongly convex.

Smoothness of Hessians We also require the Hessian of the function f to be smooth. We say a function $f(w)$ has ρ-Lipschitz Hessian if for any two points w_1, w_2 we have

$$\|\nabla^2 f(w_1) - \nabla^2 f(w_2)\| \leq \rho \|w_1 - w_2\|. \tag{4}$$

This is a third order condition that is true if the third order derivative exists and is bounded.

2.2. Tensors decomposition

A p-th order tensor is a p-dimensional array. In this paper we will mostly consider 4-th order tensors. If $T \in \mathbb{R}^{d^4}$ is a 4-th order tensor, we use $T_{i_1,i_2,i_3,i_4}(i_1, ..., i_4 \in [d])$ to denote its (i_1, i_2, i_3, i_4)-th entry.

Tensors can be constructed from tensor products. We use $(u \otimes v)$ to denote a 2nd order tensor where $(u \otimes v)_{i,j} = u_i v_j$. This generalizes to higher order and we use $u^{\otimes 4}$ to denote the 4-th order tensor

$$[u^{\otimes 4}]_{i_1,i_2,i_3,i_4} = u_{i_1} u_{i_2} u_{i_3} u_{i_4}.$$

We say a 4-th order tensor $T \in \mathbb{R}^{d^4}$ has an *orthogonal decomposition* if it can be written as

$$T = \sum_{i=1}^{d} a_i^{\otimes 4}, \tag{5}$$

where a_i's are orthonormal vectors that satisfy $\|a_i\| = 1$ and $a_i^T a_j = 0$ for $i \neq j$. We call the vectors a_i's the components of this decomposition. Such a decomposition is unique up to permutation of a_i's and sign-flips.

A tensor also defines a multilinear form (just as a matrix defines a bilinear form), for a p-th order tensor $T \in \mathbb{R}^{d^p}$ and matrices $M_i \in \mathbb{R}^{d \times n_i}, i \in [p]$, we define

$$[T(M_1, M_2, ..., M_p)]_{i_1,i_2,...,i_p} = \sum_{j_1,j_2,...,j_p \in [d]} T_{j_1,j_2,...,j_p} \prod_{t \in [p]} M_t[j_t, i_t].$$

That is, the result of the multilinear form $T(M_1, M_2, ..., M_p)$ is another tensor in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_p}$. We will most often use vectors or identity matrices in the multilinear form. In particular, for a 4-th order tensor $T \in \mathbb{R}^{d^4}$ we know $T(I, u, u, u)$ is a vector and $T(I, I, u, u)$ is a matrix. In particular, if T has the orthogonal decomposition in (5), we know $T(I, u, u, u) = \sum_{i=1}^d (u^T a_i)^3 a_i$ and $T(I, I, u, u) = \sum_{i=1}^d (u^T a_i)^2 a_i a_i^T$.

Given a tensor T with an orthogonal decomposition, the orthogonal tensor decomposition problem asks to find the individual components $a_1, ..., a_d$. This is a central problem in learning many latent variable models, including Hidden Markov Model, multi-view models, topic models, mixture of Gaussians and Independent Component Analysis (ICA). See the discussion and citations in Anandkumar et al. (2014). Orthogonal tensor decomposition problem can be solved by many algorithms even when the input is a noisy estimation $\tilde{T} \approx T$ (Harshman, 1970; Kolda, 2001; Anandkumar et al., 2014). In practice this approach has been successfully applied to ICA (Comon, 2002), topic models (Zou et al., 2013) and community detection (Huang et al., 2013).

3. Stochastic gradient descent for strict saddle function

In this section we discuss the properties of saddle points, and show if all the saddle points are well-behaved then stochastic gradient descent finds a local minimum for a non-convex function in polynomial time.

3.1. Strict saddle property

For a twice differentiable function $f(w)$, we call a point *stationary point* if its gradient is equal to 0. Stationary points could be local minima, local maxima or saddle points. By local optimality conditions (Wright and Nocedal, 1999), in many cases we can tell what type a point w is by looking at its Hessian: if $\nabla^2 f(w)$ is positive definite then w is a local minimum; if $\nabla^2 f(w)$ is negative definite then w is a local maximum; if $\nabla^2 f(w)$ has both positive and negative eigenvalues then w is a saddle point. These criteria do not cover all the cases as there could be degenerate scenarios: $\nabla^2 f(w)$ can be positive semidefinite with an eigenvalue equal to 0, in which case the point could be a local minimum or a saddle point.

If a function does not have these degenerate cases, then we say the function is strict saddle:

Definition 4 *A twice differentiable function $f(w)$ is* strict saddle*, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.*

Intuitively, if we are not at a stationary point, then we can always follow the gradient and reduce the value of the function. If we are at a saddle point, we need to consider a second order Taylor expansion:

$$f(w + \Delta w) \approx w + (\Delta w)^T \nabla^2 f(w)(\Delta w) + O(\|\Delta w\|^3).$$

Since the strict saddle property guarantees $\nabla^2 f(w)$ to have a negative eigenvalue, there is always a point that is near w and has strictly smaller function value. It is possible to make local improvements as long as we have access to second order information. However it is not clear whether the more efficient stochastic gradient updates can work in this setting.

To make sure the local improvements are significant, we use a robust version of the strict saddle property:

Definition 5 *A twice differentiable function $f(w)$ is $(\alpha, \gamma, \epsilon, \delta)$-*strict saddle*, if for any point w at least one of the following is true*

1. $\|\nabla f(w)\| \geq \epsilon$.
2. $\lambda_{min}(\nabla^2 f(w)) \leq -\gamma$.
3. *There is a local minimum $w^\star$ such that $\|w - w^\star\| \leq \delta$, and the function $f(w')$ restricted to 2δ neighborhood of $w^\star$ ($\|w' - w^\star\| \leq 2\delta$) is α-strongly convex.*

Intuitively, this condition says for any point whose gradient is small, it is either close to a robust local minimum, or is a saddle point (or local maximum) with a significant negative eigenvalue.

Algorithm 1 Noisy Stochastic Gradient

Require: Stochastic gradient oracle $SG(w)$, initial point w_0, desired accuracy κ.
Ensure: w_t that is close to some local minimum $w^\star$.

```
Choose η = min{Õ(κ²/log(1/κ)), η_max}
for t = 0 to Õ(1/η²) do
    Sample noise n uniformly from unit sphere.
    w_{t+1} ← w_t − η(SG(w) + n)
```

We purpose a simple variant of stochastic gradient algorithm, where the only difference to the traditional algorithm is we add an extra noise term to the updates. The main benefit of this additional noise is that we can guarantee there is noise in every direction, which allows the algorithm to effectively explore the local neighborhood around saddle points. If the noise from stochastic gradient oracle already has nonnegligible variance in every direction, our analysis also applies without adding additional noise. We show noise can help the algorithm escape from saddle points and optimize strict saddle functions.

Theorem 6 (Main Theorem) *Suppose a function $f(w) : \mathbb{R}^d \to \mathbb{R}$ that is $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, and has a stochastic gradient oracle with radius at most Q. Further, suppose the function is bounded by $|f(w)| \leq B$, is β-smooth and has ρ-Lipschitz Hessian. Then there exists a threshold $\eta_{\max} = \tilde{\Theta}(1)$, so that for any $\zeta > 0$, and for any $\eta \leq \eta_{\max}/\max\{1, \log(1/\zeta)\}$, with probability at least $1-\zeta$ in $t = \tilde{O}(\eta^{-2}\log(1/\zeta))$ iterations, Algorithm 1 (Noisy Gradient Descent) outputs a point w_t that is $\tilde{O}(\sqrt{\eta\log(1/\eta\zeta)})$-close to some local minimum $w^\star$.*

Here (and throughout the rest of the paper) $\tilde{O}(\cdot)$ $(\tilde{\Omega}, \tilde{\Theta})$ hides the factor that is polynomially dependent on all other parameters (including Q, $1/\alpha$, $1/\gamma$, $1/\epsilon$, $1/\delta$, B, β, ρ, and d), but independent of η and ζ. So it focuses on the dependency on η and ζ. Our proof technique can give explicit dependencies on these parameters however we hide these dependencies for simplicity of presentation. [2]

2. Currently, our number of iteration is a large polynomial in the dimension d. We have not tried to optimize the degree of this polynomial. Empirically the dependency on d is much better, whether the dependency on d can be improved to poly $\log d$ is left as an open problem.

Remark 7 (Decreasing learning rate) *Often analysis of stochastic gradient descent uses decreasing learning rates and the algorithm converges to a local (or global) minimum. Since the function is strongly convex in the small region close to local minimum, we can use Theorem 6 to first find a point that is close to a local minimum, and then apply standard analysis of SGD in the strongly convex case (where we decrease the learning rate by $1/t$ and get $1/\sqrt{t}$ convergence in $\|w - w^\star\|$).*

In the next part we sketch the proof of the main theorem. Details are deferred to Appendix A.

3.2. Proof sketch

In order to prove Theorem 6, we analyze the three cases in Definition 5. When the gradient is large, we show the function value decreases in one step (see Lemma 8); when the point is close to a local minimum, we show with high probability it cannot escape in the next polynomial number of iterations (see Lemma 9).

Lemma 8 (Gradient) *Under the assumptions of Theorem 6, for any point with $\|\nabla f(w_t)\| \geq C\sqrt{\eta}$ (where $C = \tilde{\Theta}(1)$) and $C\sqrt{\eta} \leq \epsilon$, after one iteration we have $\mathbb{E}[f(w_{t+1})] \leq f(w_t) - \tilde{\Omega}(\eta^2)$.*

The proof of this lemma is a simple application of the smoothness property.

Lemma 9 (Local minimum) *Under the assumptions of Theorem 6, for any point w_t that is $\tilde{O}(\sqrt{\eta}) < \delta$ close to local minimum $w^\star$, in $\tilde{O}(\eta^{-2}\log(1/\zeta))$ number of steps all future w_{t+i}'s are $\tilde{O}(\sqrt{\eta\log(1/\eta\zeta)})$-close with probability at least $1 - \zeta/2$.*

The proof of this lemma is similar to the standard analysis (Rakhlin et al., 2012) of stochastic gradient descent in the smooth and strongly convex setting, except we only have local strong convexity. The proof appears in Appendix A.

The hardest case is when the point is "close" to a saddle point: it has gradient smaller than ϵ and smallest eigenvalue of the Hessian bounded by $-\gamma$. In this case we show the noise in our algorithm helps the algorithm to escape:

Lemma 10 (Saddle point) *Under the assumptions of Theorem 6, for any point w_t where $\|\nabla f(w_t)\| \leq C\sqrt{\eta}$ (for the same C as in Lemma 8), and $\lambda_{\min}(\nabla^2 f(w_t)) \leq -\gamma$, there is a number of steps T that depends on w_t such that $\mathbb{E}[f(w_{t+T})] \leq f(w_t) - \tilde{\Omega}(\eta)$. The number of steps T has a fixed upper bound T_{max} that is independent of w_t where $T \leq T_{max} = \tilde{O}(1/\eta)$.*

Intuitively, at point w_t there is a good direction that is hiding in the Hessian. The hope of the algorithm is that the additional (or inherent) noise in the update step makes a small step towards the correct direction, and then the gradient information will reinforce this small perturbation and the future updates will "slide" down the correct direction.

To make this more formal, we consider a coupled sequence of updates $\tilde{w}$ such that the function to minimize is just the local second order approximation

$$\tilde{f}(w) = f(w_t) + \nabla f(w_t)^T(w - w_t) + \frac{1}{2}(w - w_t)^T \nabla^2 f(w_t)(w - w_t).$$

The dynamics of stochastic gradient descent for this quadratic function is easy to analyze as $\tilde{w}_{t+i}$ can be calculated analytically. Indeed, we show the expectation of $\tilde{f}(\tilde{w})$ will decrease. More concretely we show the point $\tilde{w}_{t+i}$ will move substantially in the negative curvature directions and remain close to w_t in positive curvature directions. We then use the smoothness of the function to show that as long as the points did not go very far from w_t, the two update sequences $\tilde{w}$ and w will remain close to each other, and thus $\tilde{f}(\tilde{w}_{t+i}) \approx f(w_{t+i})$. Finally we prove the future w_{t+i}'s (in the next T steps) will remain close to w_t with high probability by Martingale bounds. The detailed proof appears in Appendix A.

With these three lemmas it is easy to prove the main theorem. Intuitively, as long as there is a small probability of being $\tilde{O}(\sqrt{\eta})$-close to a local minimum, we can always apply Lemma 8 or Lemma 10 to make the expected function value decrease by $\tilde{\Omega}(\eta)$ in at most $\tilde{O}(1/\eta)$ iterations, this cannot go on for more than $\tilde{O}(1/\eta^2)$ iterations because in that case the expected function value will decrease by more than $2B$, but $\max f(x) - \min f(x) \leq 2B$ by our assumption. Therefore in $\tilde{O}(1/\eta^2)$ steps with at least constant probability w_t will become $\tilde{O}(\sqrt{\eta})$-close to a local minimum. By Lemma 9 we know once it is close it will almost always stay close, so after q epochs of $\tilde{O}(1/\eta^2)$ iterations each, the probability of success will be $1 - \exp(-\Omega(q))$. Taking $q = O(\log(1/\zeta))$ gives the result. More details appear in Appendix A.

3.3. Constrained Problems

In many cases, the problem we are facing are constrained optimization problems. In this part we briefly describe how to adapt the analysis to problems with equality constraints (which suffices for the tensor application). Dealing with general inequality constraint is left as future work.

For a constrained optimization problem:

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & f(w) \\ \text{s.t.} \quad & c_i(w) = 0, \qquad i \in [m] \end{aligned} \tag{6}$$

in general we need to consider the set of points in a low dimensional manifold that is defined by the constraints. In particular, in the algorithm after every step we need to project back to this manifold (see Algorithm 2 where $\Pi_{\mathcal{W}}$ is the projection to this manifold).

Algorithm 2 Projected Noisy Stochastic Gradient

Require: Stochastic gradient oracle $SG(w)$, initial point w_0, desired accuracy κ.
Ensure: w_t that is close to some local minimum $w^\star$.

1: Choose $\eta = \min\{\tilde{O}(\kappa^2/\log(1/\kappa)), \eta_{\max}\}$
2: **for** $t = 0$ to $\tilde{O}(1/\eta^2)$ **do**
3: Sample noise n uniformly from unit sphere.
4: $v_{t+1} \leftarrow w_t - \eta(SG(w) + n)$
5: $w_{t+1} = \Pi_{\mathcal{W}}(v_{t+1})$

For constrained optimization it is common to consider the Lagrangian:

$$\mathcal{L}(w, \lambda) = f(w) - \sum_{i=1}^{m} \lambda_i c_i(w). \tag{7}$$

Under common regularity conditions, it is possible to compute the value of the Lagrangian multipliers:

$$\lambda^*(w) = \arg\min_{\lambda} \|\nabla_w \mathcal{L}(w, \lambda)\|.$$

We can also define the tangent space, which contains all directions that are orthogonal to all the gradients of the constraints: $\mathcal{T}(w) = \{v : \nabla c_i(w)^T v = 0;\ i = 1, \cdots, m\}$. In this case the corresponding gradient and Hessian we consider are the first-order and second-order partial derivative of Lagrangian $\mathcal{L}$ at point $(w, \lambda^*(w))$:

$$\chi(w) = \nabla_w \mathcal{L}(w, \lambda)|_{(w,\lambda^*(w))} = \nabla f(w) - \sum_{i=1}^{m} \lambda_i^*(w) \nabla c_i(w) \tag{8}$$

$$\mathfrak{M}(w) = \nabla^2_{ww} \mathcal{L}(w, \lambda)|_{(w,\lambda^*(w))} = \nabla^2 f(w) - \sum_{i=1}^{m} \lambda_i^*(w) \nabla^2 c_i(w) \tag{9}$$

We replace the gradient and Hessian with $\chi(w)$ and $\mathfrak{M}(w)$, and when computing eigenvectors of $\mathfrak{M}(w)$ we focus on its projection on the tangent space. In this way, we can get a similar definition for strict saddle (see Appendix B), and the following theorem.

Theorem 11 *(informal) Under regularity conditions and smoothness conditions, if a constrained optimization problem satisfies strict saddle property, then for a small enough η, in $\tilde{O}(\eta^{-2} \log 1/\zeta)$ iterations Projected Noisy Gradient Descent (Algorithm 2) outputs a point w that is $\tilde{O}(\sqrt{\eta} \log(1/\eta\zeta))$ close to a local minimum with probability at least $1 - \zeta$.*

Detailed discussions and formal version of this theorem are deferred to Appendix B.

4. Online Tensor Decomposition

In this section we describe how to apply our stochastic gradient descent analysis to tensor decomposition problems. We first give a new formulation of tensor decomposition as an optimization problem, and show that it satisfies the strict saddle property. Then we explain how to compute stochastic gradient in a simple example of Independent Component Analysis (ICA) (Hyvärinen et al., 2004).

4.1. Optimization problem for tensor decomposition

Given a tensor $T \in \mathbb{R}^{d^4}$ that has an orthogonal decomposition

$$T = \sum_{i=1}^{d} a_i^{\otimes 4}, \tag{10}$$

where the components a_i's are orthonormal vectors ($\|a_i\| = 1$, $a_i^T a_j = 0$ for $i \neq j$), the goal of orthogonal tensor decomposition is to find the components a_i's. This problem has inherent symmetry: for any permutation π and any set of $\kappa_i \in \{\pm 1\}, i \in [d]$, we know $u_i = \kappa_i a_{\pi(i)}$ is also a valid solution. This symmetry property makes the natural optimization problems non-convex.

In this section we will give a new formulation of orthogonal tensor decomposition as an optimization problem, and show that this new problem satisfies the strict saddle property. Previously, Frieze et al. (1996) solves the problem of finding one component, with the following objective function

$$\max_{\|u\|^2=1} \quad T(u, u, u, u). \tag{11}$$

In Appendix C.1, as a warm-up example we show this function is indeed strict saddle, and we can apply Theorem 11 to prove global convergence of stochastic gradient descent algorithm.

It is possible to find all components of a tensor by iteratively finding one component, and do careful *deflation*, as described in Anandkumar et al. (2014) or Arora et al. (2012). However, in practice the most popular approaches like Alternating Least Squares (Comon et al., 2009) or FastICA (Hyvarinen, 1999) try to use a single optimization problem to find all the components. Empirically these algorithms are often more robust to noise and model misspecification.

The most straight-forward formulation of the problem aims to minimize the *reconstruction error*

$$\min_{\forall i, \|u_i\|^2=1} \quad \|T - \sum_{i=1}^{d} u_i^{\otimes 4}\|_F^2. \tag{12}$$

Here $\|\cdot\|_F$ is the Frobenius norm of the tensor which is equal to the ℓ_2 norm when we view the tensor as a d^4 dimensional vector. However, it is not clear whether this function satisfies the strict saddle property, and empirically stochastic gradient descent is unstable for this objective.

We propose a new objective that aims to minimize the correlation between different components:

$$\min_{\forall i, \|u_i\|^2=1} \quad \sum_{i \neq j} T(u_i, u_i, u_j, u_j), \tag{13}$$

To understand this objective intuitively, we first expand vectors u_k in the orthogonal basis formed by $\{a_i\}$'s. That is, we can write $u_k = \sum_{i=1}^{d} z_k(i) a_i$, where $z_k(i)$ are scalars that correspond to the coordinates in the $\{a_i\}$ basis. In this way we can rewrite $T(u_k, u_k, u_l, u_l) = \sum_{i=1}^{d} (z_k(i))^2 (z_l(i))^2$. From this form it is clear that the $T(u_k, u_k, u_l, u_l)$ is always nonnegative, and is equal to 0 only when the support of z_k and z_l do not intersect. For the objective function, we know in order for it to be equal to 0 the z's must have disjoint support. Therefore, we claim that $\{u_k\}, \forall k \in [d]$ is equivalent to $\{a_i\}, \forall i \in [d]$ up to permutation and sign flips when the global minimum (which is 0) is achieved.

We further show that this optimization program satisfies the strict saddle property and all its local minima in fact achieves global minimum value. The proof is deferred to Appendix C.2.

Theorem 12 *The optimization problem (13) is $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, for $\alpha = 1$ and $\gamma, \epsilon, \delta = 1/poly(d)$. Moreover, all its local minima have the form $u_i = \kappa_i a_{\pi(i)}$ for some $\kappa_i = \pm 1$ and permutation $\pi(i)$.*

Note that we can also generalize this to handle 4th order tensors with different positive weights on the components, or other order tensors, see Appendix C.3.

4.2. Implementing stochastic gradient oracle

To design an online algorithm based on objective function (13), we need to give an implementation for the stochastic gradient oracle.

In applications, the tensor T is oftentimes the expectation of multilinear operations of samples $g(x)$ over x where x is generated from some distribution $\mathcal{D}$. In other words, for any $x \sim \mathcal{D}$, the tensor is $T = \mathbb{E}[g(x)]$. Using the linearity of the multilinear map, we know $\mathbb{E}[g(x)](u_i, u_i, u_j, u_j) = \mathbb{E}[g(x)(u_i, u_i, u_j, u_j)]$. Therefore we can define the loss function $\phi(u, x) = \sum_{i \neq j} g(x)(u_i, u_i, u_j, u_j)$, and the stochastic gradient oracle $SG(u) = \nabla_u \phi(u, x)$.

For concreteness, we look at a simple ICA example. In the simple setting we consider an unknown signal x that is uniform[3] in $\{\pm 1\}^d$, and an unknown orthonormal linear transformation[4] A ($AA^T = I$). The sample we observe is $y := Ax \in \mathbb{R}^d$. Using standard techniques (see Cardoso (1989)), we know the 4-th order cumulant of the observed sample is a tensor that has orthogonal decomposition. Here for simplicity we don't define 4-th order cumulant, instead we give the result directly.

3. In general ICA the entries of x are independent, non-Gaussian variables.
4. In general (under-complete) ICA this could be an arbitrary linear transformation, however usually after the "whitening" step (see Cardoso (1989)) the linear transformation becomes orthonormal.

Define tensor $Z \in \mathbb{R}^{d^4}$ as follows:

$$\begin{aligned} &Z(i,i,i,i) = 3, && \forall i \in [d] \\ &Z(i,i,j,j) = Z(i,j,i,j) = Z(i,j,j,i) = 1, && \forall i \neq j \in [d] \end{aligned}$$

where all other entries of Z are equal to 0. The tensor T can be written as a function of the auxiliary tensor Z and multilinear form of the sample y.

Lemma 13 *The expectation $\mathbb{E}[\frac{1}{2}(Z - y^{\otimes 4})] = \sum_{i=1}^d a_i^{\otimes 4} = T$, where a_i's are columns of the unknown orthonormal matrix A.*

This lemma is easy to verify, and is closely related to cumulants (Cardoso, 1989). Recall that $\phi(u,y)$ denotes the loss (objective) function evaluated at sample y for point u. Let $\phi(u,y) = \sum_{i\neq j} \frac{1}{2}(Z - y^{\otimes 4})(u_i, u_i, u_j, u_j)$. By Lemma 13, we know that $\mathbb{E}[\phi(u,y)]$ is equal to the objective function as in Equation (13). Therefore we rewrite objective (13) as the following stochastic optimization problem

$$\min_{\forall i, \|u_i\|^2=1} \quad \mathbb{E}[\phi(u,y)], \text{ where } \phi(u,y) = \sum_{i\neq j} \frac{1}{2}(Z - y^{\otimes 4})(u_i, u_i, u_j, u_j)$$

The stochastic gradient oracle is then

$$\nabla_{u_i}\phi(u,y) = \sum_{j\neq i} \Big(\langle u_j, u_j \rangle u_i + 2 \langle u_i, u_j \rangle u_j - \langle u_j, y \rangle^2 \langle u_i, y \rangle y \Big). \tag{14}$$

Notice that computing this stochastic gradient does not require constructing the 4-th order tensor $T - y^{\otimes 4}$. In particular, this stochastic gradient can be computed very efficiently:

Remark *The stochastic gradient (14) can be computed for all u_i's in $O(d^3)$ time for one sample or $O(d^3 + d^2 k)$ for average of k samples.*

Proof The proof is straight forward as the first two terms on the right hand side take $O(d^3)$ and is shared by all samples. The third term can be efficiently computed once the inner-products between all the y's and all the u_i's are computed (which takes $O(kd^2)$ time). ∎

5. Experiments

We run simulations for Projected Noisy Gradient Descent (Algorithm 2) applied to orthogonal tensor decomposition. The results show that the algorithm converges from random initial points efficiently (as predicted by the theorems), and our new formulation (13) performs better than reconstruction error (12) based formulation.

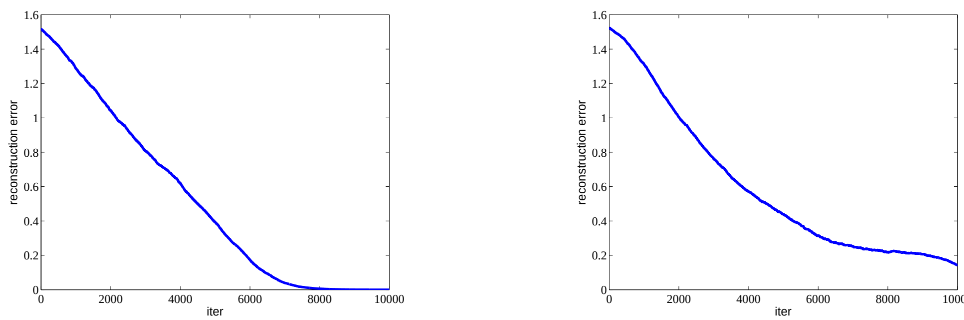
Settings We set dimension $d = 10$, the input tensor T is a random tensor in $\mathbb{R}^{10^4}$ that has orthogonal decomposition (5). The step size is chosen carefully for respective objective functions. The performance is measured by normalized reconstruction error $\mathcal{E} = \left(\|T - \sum_{i=1}^d u_i^{\otimes 4}\|_F^2\right) / \|T\|_F^2$.

Samples and stochastic gradients We use two ways to generate samples and compute stochastic gradients. In the first case we generate sample x by setting it equivalent to $d^{\frac{1}{4}} a_i$ with probability $1/d$. It is easy to see that $\mathbb{E}[x^{\otimes 4}] = T$. This is a very simple way of generating samples, and we use it as a sanity check for the objective functions.

In the second case we consider the ICA example introduced in Section 4.2, and use Equation (14) to compute a stochastic gradient. In this case the stochastic gradient has a large variance, so we use mini-batch of size 100 to reduce the variance.

Comparison of objective functions We use the simple way of generating samples for our new objective function (13) and reconstruction error objective (12). The result is shown in Figure 1. Our new objective function is empirically more stable (always converges within 10000 iterations); the reconstruction error do not always converge within the same number of iterations and often exhibits long periods with small improvement (which is likely to be caused by saddle points that do not have a significant negative eigenvalue).

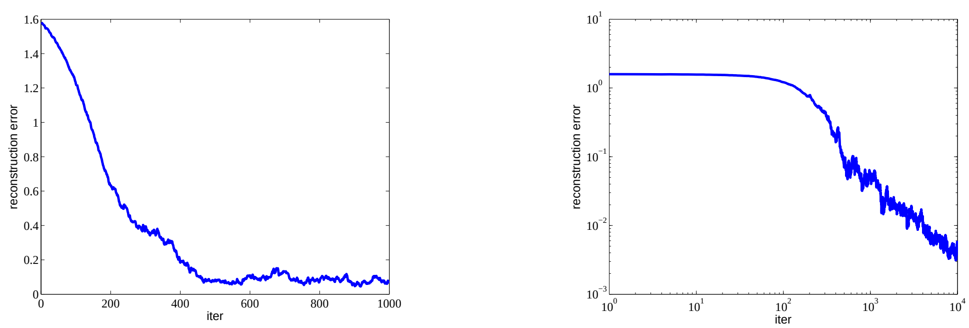
Simple ICA example As shown in Figure 2, our new algorithm also works in the ICA setting. When the learning rate is constant the error stays at a fixed small value. When we decrease the learning rate the error converges to 0.



(a) New Objective (13)

(b) Reconstruction Error Objective (12)

Figure 1: Comparison of different objective functions



(a) Constant Learning Rate η

(b) Learning Rate η/t (in $\log$ scale)

Figure 2: ICA setting performance with mini-batch of size 100

6. Conclusion

In this paper we identify the strict saddle property and show stochastic gradient descent converges to a local minimum under this assumption. This leads to new online algorithm for orthogonal tensor decomposition. We hope this is a first step towards understanding stochastic gradient for more classes of non-convex functions. We believe strict saddle property can be extended to handle more functions, especially those functions that have similar symmetry properties.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1): 1–127, 2009.

J-F Cardoso. Source separation using higher order moments. In *Acoustics, Speech, and Signal Processing*, pages 2109–2112. IEEE, 1989.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. *arXiv:1412.0233*, 2014.

P. Comon. Tensor decompositions. *Mathematics in Signal Processing V*, pages 1–24, 2002.

Pierre Comon, Xavier Luciani, and André LF De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7-8):393–405, 2009.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 359–359, 1996.

Morgan A Hanson. Invexity and the kuhn–tucker theorem. *Journal of mathematical analysis and applications*, 236(2):594–604, 1999.

Richard A Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16(1):84, 1970.

Furong Huang, UN Niranjan, Mohammad Umar Hakeem, and Animashree Anandkumar. Fast detection of overlapping communities via online tensor methods. *arXiv:1309.0787*, 2013.

Aapo Hyvarinen. Fast ICA for noisy data using gaussian moments. In *Circuits and Systems*, volume 5, pages 57–61, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Masato Inoue, Hyeyoung Park, and Masato Okada. On-line learning theory of soft committee machines with correlated hidden units–steepest gradient descent and natural gradient descent–. *Journal of the Physical Society of Japan*, 72(4):805–810, 2003.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

David S Johnson, Christos H Papadimitriou, and Mihalis Yannakakis. How easy is local search? *Journal of computer and system sciences*, 37(1):79–100, 1988.

Krzysztof C Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90(1):1–25, 2001.

Tamara G Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.

Olvi L Mangasarian. Pseudo-convex functions. *Journal of the Society for Industrial & Applied Mathematics, Series A: Control*, 3(2):281–290, 1965.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, pages 449–456, 2012.

Magnus Rattray, David Saad, and Shun-ichi Amari. Natural gradient descent for on-line learning. *Physical review letters*, 81(24):5461, 1998.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5, 1988.

David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120*, 2013.

Shai Shalev-Shwartz, Ohad Shamir, Karthik Sridharan, and Nathan Srebro. Stochastic convex optimization. In *Proceedings of The 22nd Conference on Learning Theory*, 2009.

Stephen J Wright and Jorge Nocedal. *Numerical optimization*, volume 2. Springer New York, 1999.

J. Y. Zou, D. Hsu, D. C. Parkes, and R. P. Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pages 2238–2246, 2013.

Appendix A. Detailed Analysis for Section 3 in Unconstrained Case

In this section we give detailed analysis for noisy gradient descent, under the assumption that the unconstrained problem satisfies $(\alpha, \gamma, \epsilon, \delta)$-strict saddle property.

The algorithm we investigate in Algorithm 1, we can combine the randomness in the stochastic gradient oracle and the artificial noise, and rewrite the update equation in form:

$$w_t = w_{t-1} - \eta(\nabla f(w_{t-1}) + \xi_{t-1}) \tag{15}$$

where η is step size, $\xi = SG(w_{t-1}) - \nabla f(w_{t-1}) + n$ (recall n is a random vector on unit sphere) is the combination of two source of noise.

By assumption, we know ξ's are independent and they satisfying $\mathbb{E}\xi = 0$, $\|\xi\| \le Q + 1$. Due to the explicitly added noise in Algorithm 1, we further have $\mathbb{E}\xi\xi^T \succ \frac{1}{d}I$. For simplicity, we assume $\mathbb{E}\xi\xi^T = \sigma^2 I$, for some constant $\sigma = \tilde{\Theta}(1)$, then the algorithm we are running is exactly the same as Stochastic Gradient Descent (SGD). Our proof can be very easily extended to the case when $\frac{1}{d}I \preceq \mathbb{E}[\xi\xi^T] \preceq (Q + \frac{1}{d})I$ because both the upper and lower bounds are $\tilde{\Theta}(1)$.

We first restate the main theorem in the context of stochastic gradient descent.

Theorem 14 (Main Theorem) *Suppose a function $f(w) : \mathbb{R}^d \to \mathbb{R}$ that is $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, and has a stochastic gradient oracle where the noise satisfy $\mathbb{E}\xi\xi^T = \sigma^2 I$. Further, suppose the function is bounded by $|f(w)| \le B$, is β-smooth and has ρ-Lipschitz Hessian. Then there exists a threshold $\eta_{\max} = \tilde{\Theta}(1)$, so that for any $\zeta > 0$, and for any $\eta \le \eta_{\max}/\max\{1, \log(1/\zeta)\}$, with probability at least $1 - \zeta$ in $t = \tilde{O}(\eta^{-2}\log(1/\zeta))$ iterations, SGD outputs a point w_t that is $\tilde{O}(\sqrt{\eta\log(1/\eta\zeta)})$-close to some local minimum $w^\star$.*

Recall that $\tilde{O}(\cdot)$ $(\tilde{\Omega}, \tilde{\Theta})$ hides the factor that has polynomial dependence on all other parameters, but is independent of η and ζ. So it focuses on the dependency on η and ζ. Throughout the proof, we interchangeably use both $\mathcal{H}(w)$ and $\nabla^2 f(w)$ to represent the Hessian matrix of $f(w)$.

As we discussed in the proof sketch in Section 3, we analyze the behavior of the algorithm in three different cases. The first case is when the gradient is large.

Lemma 15 *Under the assumptions of Theorem 14, for any point with $\|\nabla f(w_0)\| \ge \sqrt{2\eta\sigma^2\beta d}$ where $\sqrt{2\eta\sigma^2\beta d} < \epsilon$, after one iteration we have:*

$$\mathbb{E}f(w_1) - f(w_0) \le -\tilde{\Omega}(\eta^2) \tag{16}$$

Proof Our assumption can guarantee $\eta_{\max} < \frac{1}{\beta}$, then by update equation Eq.(15), we have:

$$\begin{aligned}
\mathbb{E}f(w_1) - f(w_0) &\le \nabla f(w_0)^T\mathbb{E}(w_1 - w_0) + \frac{\beta}{2}\mathbb{E}\|w_1 - w_0\|^2 \\
&= \nabla f(w_0)^T\mathbb{E}\left(-\eta(\nabla f(w_0) + \xi_0)\right) + \frac{\beta}{2}\mathbb{E}\left\|-\eta(\nabla f(w_0) + \xi_0)\right\|^2 \\
&= -(\eta - \frac{\beta\eta^2}{2})\|\nabla f(w_0)\|^2 + \frac{\eta^2\sigma^2\beta d}{2} \\
&\le -\frac{\eta}{2}\|\nabla f(w_0)\|^2 + \frac{\eta^2\sigma^2\beta d}{2} \le -\frac{\eta^2\sigma^2\beta d}{2}
\end{aligned} \tag{17}$$

which finishes the proof. ∎

Lemma 16 *Under the assumptions of Theorem 14, for any initial point w_0 that is $\tilde{O}(\sqrt{\eta}) < \delta$ close to a local minimum $w^\star$, with probability at least $1 - \zeta/2$, we have following holds simultaneously:*

$$\forall t \le \tilde{O}(\frac{1}{\eta^2}\log\frac{1}{\zeta}), \quad \|w_t - w^\star\| \le \tilde{O}(\sqrt{\eta \log\frac{1}{\eta\zeta}}) < \delta \tag{18}$$

where $w^\star$ is the locally optimal point.

Proof We shall construct a supermartingale and use Azuma's inequality (Azuma, 1967) to prove this result.

Let filtration $\mathfrak{F}_t = \sigma\{\xi_0, \cdots \xi_{t-1}\}$, and note $\sigma\{\Delta_0, \cdots, \Delta_t\} \subset \mathfrak{F}_t$, where $\sigma\{\cdot\}$ denotes the sigma field. Let event $\mathfrak{E}_t = \{\forall \tau \le t, \|w_\tau - w^\star\| \le \mu\sqrt{\eta \log\frac{1}{\eta\zeta}} < \delta\}$, where μ is independent of (η, ζ), and will be specified later. To ensure the correctness of proof, $\tilde{O}$ notation in this proof will never hide any dependence on μ. Clearly there's always a small enough choice of $\eta_{\max} = \tilde{\Theta}(1)$ to make $\mu\sqrt{\eta\log\frac{1}{\eta\zeta}} < \delta$ holds as long as $\eta \le \eta_{\max}/\max\{1, \log(1/\zeta)\}$. Also note $\mathfrak{E}_t \subset \mathfrak{E}_{t-1}$, that is $1_{\mathfrak{E}_t} \le 1_{\mathfrak{E}_{t-1}}$.

By Definition 5 of $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, we know f is locally α-strongly convex in the 2δ-neighborhood of $w^\star$. Since $\nabla f(w^\star) = 0$, we have

$$\nabla f(w_t)^T(w_t - w^\star)1_{\mathfrak{E}_t} \ge \alpha\|w_t - w^\star\|^2 1_{\mathfrak{E}_t} \tag{19}$$

Furthermore, with $\eta_{\max} < \frac{\alpha}{\beta^2}$, using β-smoothness, we have:

$$\begin{aligned}
\mathbb{E}[\|w_t - w^\star\|^2 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}] =& \mathbb{E}[\|w_{t-1} - \eta(\nabla f(w_{t-1}) + \xi_{t-1}) - w^\star\|^2 | \mathfrak{F}_{t-1}] 1_{\mathfrak{E}_{t-1}} \\
=& \left[\|w_{t-1} - w^\star\|^2 - 2\eta\nabla f(w_{t-1})^T(w_{t-1} - w^\star) + \eta^2\|\nabla f(w_{t-1})\|^2 + \eta^2 d\sigma^2\right] 1_{\mathfrak{E}_{t-1}} \\
\le& [(1 - 2\eta\alpha + \eta^2\beta^2)\|w_{t-1} - w^\star\|^2 + \eta^2 d\sigma^2] 1_{\mathfrak{E}_{t-1}} \\
\le& [(1 - \eta\alpha)\|w_{t-1} - w^\star\|^2 + \eta^2 d\sigma^2] 1_{\mathfrak{E}_{t-1}}
\end{aligned} \tag{20}$$

Therefore, we have:

$$\left[\mathbb{E}[\|w_t - w^\star\|^2 | \mathfrak{F}_{t-1}] - \frac{\eta d\sigma^2}{\alpha}\right] 1_{\mathfrak{E}_{t-1}} \le (1 - \eta\alpha)\left[\|w_{t-1} - w^\star\|^2 - \frac{\eta d\sigma^2}{\alpha}\right] 1_{\mathfrak{E}_{t-1}} \tag{21}$$

Then, let $G_t = \max\{(1-\eta\alpha)^{-t}(\|w_t - w^\star\|^2 - \frac{\eta d\sigma^2}{\alpha}), 0\}$, we have:

$$\mathbb{E}[G_t 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}] \le G_{t-1}1_{\mathfrak{E}_{t-1}} \le G_{t-1}1_{\mathfrak{E}_{t-2}} \tag{22}$$

which means $G_t 1_{\mathfrak{E}_{t-1}}$ is a supermartingale.

Therefore, with probability 1, we have:

$$\begin{aligned}
&|G_t 1_{\mathfrak{E}_{t-1}} - \mathbb{E}[G_t 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}]| \\
\le& (1-\eta\alpha)^{-t}[\ \|w_{t-1} - \eta\nabla f(w_{t-1}) - w^\star\| \cdot \eta\|\xi_{t-1}\| + \eta^2\|\xi_{t-1}\|^2 + \eta^2 d\sigma^2\ ] 1_{\mathfrak{E}_{t-1}} \\
\le& (1-\eta\alpha)^{-t} \cdot \tilde{O}(\mu\eta^{1.5}\log^{\frac{1}{2}}\frac{1}{\eta\zeta}) = d_t
\end{aligned} \tag{23}$$

Let

$$c_t = \sqrt{\sum_{\tau=1}^{t} d_\tau^2} = \tilde{O}(\mu\eta^{1.5}\log^{\frac{1}{2}}\frac{1}{\eta\zeta})\sqrt{\sum_{\tau=1}^{t}(1-\eta\alpha)^{-2\tau}} \tag{24}$$

By Azuma's inequality, with probability less than $\tilde{O}(\eta^3\zeta)$, we have:

$$G_t 1_{\mathfrak{E}_{t-1}} > \tilde{O}(1) c_t \log^{\frac{1}{2}}(\frac{1}{\eta\zeta}) + G_0 \tag{25}$$

We know $G_t > \tilde{O}(1)c_t \log^{\frac{1}{2}}(\frac{1}{\eta\zeta}) + G_0$ is equivalent to:

$$\|w_t - w^\star\|^2 > \tilde{O}(\eta) + \tilde{O}(1)(1-\eta\alpha)^t c_t \log^{\frac{1}{2}}(\frac{1}{\eta\zeta}) \tag{26}$$

We know:

$$\begin{aligned}(1-\eta\alpha)^t c_t \log^{\frac{1}{2}}(\frac{1}{\eta\zeta}) &= \mu \cdot \tilde{O}(\eta^{1.5}\log\frac{1}{\eta\zeta})\sqrt{\sum_{\tau=1}^{t}(1-\eta\alpha)^{2(t-\tau)}} \\ =&\mu \cdot \tilde{O}(\eta^{1.5}\log\frac{1}{\eta\zeta})\sqrt{\sum_{\tau=0}^{t-1}(1-\eta\alpha)^{2\tau}} \le \mu \cdot \tilde{O}(\eta^{1.5}\log\frac{1}{\eta\zeta})\sqrt{\frac{1}{1-(1-\eta\alpha)^2}} = \mu\cdot\tilde{O}(\eta\log\frac{1}{\eta\zeta})\end{aligned} \tag{27}$$

This means Azuma's inequality implies, there exist some $\tilde{C} = \tilde{O}(1)$ so that:

$$P\left(\mathfrak{E}_{t-1} \cap \left\{\|w_t - w^\star\|^2 > \mu \cdot \tilde{C}\eta\log\frac{1}{\eta\zeta})\right\}\right) \le \tilde{O}(\eta^3\zeta) \tag{28}$$

By choosing $\mu > \tilde{C}$, this is equivalent to:

$$P\left(\mathfrak{E}_{t-1} \cap \left\{\|w_t - w^\star\|^2 > \mu^2\eta\log\frac{1}{\eta\zeta}\right\}\right) \le \tilde{O}(\eta^3\zeta) \tag{29}$$

Then we have:

$$P(\overline{\mathfrak{E}}_t) = P\left(\mathfrak{E}_{t-1} \cap \left\{\|w_t - w^\star\| > \mu\sqrt{\eta\log\frac{1}{\eta\zeta}}\right\}\right) + P(\overline{\mathfrak{E}}_{t-1}) \le \tilde{O}(\eta^3\zeta) + P(\overline{\mathfrak{E}}_{t-1}) \tag{30}$$

By initialization conditions, we know $P(\overline{\mathfrak{E}}_0) = 0$, and thus $P(\overline{\mathfrak{E}}_t) \le t\tilde{O}(\eta^3\zeta)$. Take $t = \tilde{O}(\frac{1}{\eta^2}\log\frac{1}{\zeta})$, we have $P(\overline{\mathfrak{E}}_t) \le \tilde{O}(\eta\zeta\log\frac{1}{\zeta})$. When $\eta_{\max} = \tilde{O}(1)$ is chosen small enough, and $\eta \le \eta_{\max}/\log(1/\zeta)$, this finishes the proof. ■

Lemma 17 *Under the assumptions of Theorem 14, for any initial point w_0 where $\|\nabla f(w_0)\| \le \sqrt{2\eta\sigma^2\beta d} < \epsilon$, and $\lambda_{\min}(\mathcal{H}(w_0)) \le -\gamma$, then there is a number of steps T that depends on w_0 such that:*

$$\mathbb{E}f(w_T) - f(w_0) \le -\tilde{\Omega}(\eta) \tag{31}$$

The number of steps T has a fixed upper bound T_{max} that is independent of w_0 where $T \le T_{max} = O((\log d)/\gamma\eta)$.

Remark 18 *In general, if we relax the assumption $\mathbb{E}\xi\xi^T = \sigma^2 I$ to $\sigma^2_{\min} I \preceq \mathbb{E}\xi\xi^T \preceq \sigma^2_{\max} I$, the upper bound T_{max} of number of steps required in Lemma 17 would be increased to $T_{max} = O(\frac{1}{\gamma\eta}(\log d + \log\frac{\sigma_{\max}}{\sigma_{\min}}))$*

As we described in the proof sketch, the main idea is to consider a coupled update sequence that correspond to the local second-order approximation of $f(x)$ around w_0. We characterize this sequence of update in the next lemma.

Lemma 19 *Under the assumptions of Theorem 14. Let $\tilde{f}$ defined as local second-order approximation of $f(x)$ around w_0:*

$$\tilde{f}(w) \doteq f(w_0) + \nabla f(w_0)^T (w - w_0) + \frac{1}{2}(w - w_0)^T \mathcal{H}(w_0)(w - w_0) \tag{32}$$

$\{\tilde{w}_t\}$ be the corresponding sequence generated by running SGD on function $\tilde{f}$, with $\tilde{w}_0 = w_0$. For simplicity, denote $\mathcal{H} = \mathcal{H}(w_0) = \nabla^2 f(w_0)$, then we have analytically:

$$\nabla \tilde{f}(\tilde{w}_t) = (1 - \eta\mathcal{H})^t \nabla f(w_0) - \eta\mathcal{H}\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_\tau \tag{33}$$

$$\tilde{w}_t - w_0 = -\eta\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^\tau \nabla f(w_0) - \eta\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_\tau \tag{34}$$

Furthermore, for any initial point w_0 where $\|\nabla f(w_0)\| \le \tilde{O}(\eta) < \epsilon$, and $\lambda_{\min}(\mathcal{H}(w_0)) = -\gamma_0$. Then, there exist a $T \in \mathbb{N}$ satisfying:

$$\frac{d}{\eta\gamma_0} \le \sum_{\tau=0}^{T-1}(1 + \eta\gamma_0)^{2\tau} < \frac{3d}{\eta\gamma_0} \tag{35}$$

with probability at least $1 - \tilde{O}(\eta^3)$, we have following holds simultaneously for all $t \le T$:

$$\|\tilde{w}_t - w_0\| \le \tilde{O}(\eta^{\frac{1}{2}} \log\frac{1}{\eta}); \qquad \|\nabla \tilde{f}(\tilde{w}_t)\| \le \tilde{O}(\eta^{\frac{1}{2}} \log\frac{1}{\eta}) \tag{36}$$

Proof Denote $\mathcal{H} = \mathcal{H}(w_0)$, since $\tilde{f}$ is quadratic, clearly we have:

$$\nabla \tilde{f}(\tilde{w}_t) = \nabla \tilde{f}(\tilde{w}_{t-1}) + \mathcal{H}(\tilde{w}_t - \tilde{w}_{t-1}) \tag{37}$$

Substitute the update equation of SGD in Eq.(37), we have:

$$\begin{aligned}
\nabla \tilde{f}(\tilde{w}_t) &= \nabla \tilde{f}(\tilde{w}_{t-1}) - \eta\mathcal{H}(\nabla \tilde{f}(\tilde{w}_{t-1}) + \xi_{t-1}) \\
&= (1 - \eta\mathcal{H})\nabla \tilde{f}(\tilde{w}_{t-1}) - \eta\mathcal{H}\xi_{t-1} \\
&= (1 - \eta\mathcal{H})^2\nabla \tilde{f}(\tilde{w}_{t-2}) - \eta\mathcal{H}\xi_{t-1} - \eta\mathcal{H}(1 - \eta\mathcal{H})\xi_{t-2} = \cdots \\
&= (1 - \eta\mathcal{H})^t \nabla f(w_0) - \eta\mathcal{H}\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_\tau
\end{aligned} \tag{38}$$

Therefore, we have:

$$\begin{aligned}
\tilde{w}_t - w_0 &= -\eta\sum_{\tau=0}^{t-1}(\nabla \tilde{f}(\tilde{w}_\tau) + \xi_\tau) \\
&= -\eta\sum_{\tau=0}^{t-1}\left((1 - \eta\mathcal{H})^\tau \nabla f(w_0) - \eta\mathcal{H}\sum_{\tau'=0}^{\tau-1}(1 - \eta\mathcal{H})^{\tau-\tau'-1}\xi_{\tau'} + \xi_\tau\right) \\
&= -\eta\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^\tau \nabla f(w_0) - \eta\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_\tau
\end{aligned} \tag{39}$$

Next, we prove the existence of T in Eq.(35). Since $\sum_{\tau=0}^{t}(1+\eta\gamma_0)^{2\tau}$ is monotonically increasing w.r.t t, and diverge to infinity as $t\to\infty$. We know there is always some $T\in\mathbb{N}$ gives $\frac{d}{\eta\gamma_0}\le\sum_{\tau=0}^{T-1}(1+\eta\gamma_0)^{2\tau}$. Let T be the smallest integer satisfying above equation. By assumption, we know $\gamma\le\gamma_0\le L$, and

$$\sum_{\tau=0}^{t+1}(1+\eta\gamma_0)^{2\tau}=1+(1+\eta\gamma_0)^2\sum_{\tau=0}^{t}(1+\eta\gamma_0)^{2\tau} \tag{40}$$

we can choose $\eta_{\max}<\min\{(\sqrt{2}-1)/L, 2d/\gamma\}$ so that

$$\frac{d}{\eta\gamma_0}\le\sum_{\tau=0}^{T-1}(1+\eta\gamma_0)^{2\tau}\le 1+\frac{2d}{\eta\gamma_0}\le\frac{3d}{\eta\gamma_0} \tag{41}$$

Finally, by Eq.(35), we know $T=O(\log d/\gamma_0\eta)$, and $(1+\eta\gamma_0)^T\le\tilde{O}(1)$. Also because $\mathbb{E}\xi=0$ and $\|\xi\|\le Q=\tilde{O}(1)$ with probability 1, then by Hoeffding inequality, we have for each dimension i and time $t\le T$:

$$P\left(|\eta\sum_{\tau=0}^{t-1}(1-\eta\mathcal{H})^{t-\tau-1}\xi_{\tau,i}|>\tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta})\right)\le e^{-\tilde{\Omega}(\log^2\frac{1}{\eta})}\le\tilde{O}(\eta^4) \tag{42}$$

then by summing over dimension d and taking union bound over all $t\le T$, we directly have:

$$P\left(\forall t\le T, \|\eta\sum_{\tau=0}^{t-1}(1-\eta\mathcal{H})^{t-\tau-1}\xi_{\tau}\|>\tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta})\right)\le\tilde{O}(\eta^3). \tag{43}$$

Combine this fact with Eq.(38) and Eq.(39), we finish the proof.

■

Next we need to prove that the two sequences of updates are always close.

Lemma 20 *Under the assumptions of Theorem 14. and let $\{w_t\}$ be the corresponding sequence generated by running SGD on function f. Also let $\tilde{f}$ and $\{\tilde{w}_t\}$ be defined as in Lemma 19. Then, for any initial point w_0 where $\|\nabla f(w_0)\|\le\tilde{O}(\eta)<\epsilon$, and $\lambda_{\min}(\nabla^2 f(w_0))=-\gamma_0$. Given the choice of T as in Eq.(35), with probability at least $1-\tilde{O}(\eta^2)$, we have following holds simultaneously for all $t\le T$:*

$$\|w_t-\tilde{w}_t\|\le\tilde{O}(\eta\log^2\frac{1}{\eta});\qquad\|\nabla f(w_t)-\nabla\tilde{f}(\tilde{w}_t)\|\le\tilde{O}(\eta\log^2\frac{1}{\eta}) \tag{44}$$

Proof

First, we have update function of gradient by:

$$\begin{aligned}\nabla f(w_t)=&\nabla f(w_{t-1})+\int_0^1\mathcal{H}(w_{t-1}+t(w_t-w_{t-1}))\mathrm{d}t\cdot(w_t-w_{t-1})\\=&\nabla f(w_{t-1})+\mathcal{H}(w_{t-1})(w_t-w_{t-1})+\theta_{t-1}\end{aligned} \tag{45}$$

where the remainder:

$$\theta_{t-1}\equiv\int_0^1\left[\mathcal{H}(w_{t-1}+t(w_t-w_{t-1}))-\mathcal{H}(w_{t-1})\right]\mathrm{d}t\cdot(w_t-w_{t-1}) \tag{46}$$

Denote $\mathcal{H} = \mathcal{H}(w_0)$, and $\mathcal{H}'_{t-1} = \mathcal{H}(w_{t-1}) - \mathcal{H}(w_0)$. By Hessian smoothness, we immediately have:

$$\|\mathcal{H}'_{t-1}\| = \|\mathcal{H}(w_{t-1}) - \mathcal{H}(w_0)\| \le \rho\|w_{t-1} - w_0\| \le \rho(\|w_t - \tilde{w}_t\| + \|\tilde{w}_t - w_0\|) \tag{47}$$

$$\|\theta_{t-1}\| \le \frac{\rho}{2}\|w_t - w_{t-1}\|^2 \tag{48}$$

Substitute the update equation of SGD (Eq.(15)) into Eq.(45), we have:

$$\begin{aligned}\nabla f(w_t) &= \nabla f(w_{t-1}) - \eta(\mathcal{H} + \mathcal{H}'_{t-1})(\nabla f(w_{t-1}) + \xi_{t-1}) + \theta_{t-1} \\ &= (1 - \eta\mathcal{H})\nabla f(w_{t-1}) - \eta\mathcal{H}\xi_{t-1} - \eta\mathcal{H}'_{t-1}(\nabla f(w_{t-1}) + \xi_{t-1}) + \theta_{t-1}\end{aligned} \tag{49}$$

Let $\Delta_t = \nabla f(w_t) - \nabla\tilde{f}(\tilde{w}_t)$ denote the difference in gradient, then from Eq.(38), Eq.(49), and Eq.(15), we have:

$$\Delta_t = (1 - \eta\mathcal{H})\Delta_{t-1} - \eta\mathcal{H}'_{t-1}[\Delta_{t-1} + \nabla\tilde{f}(\tilde{w}_{t-1}) + \xi_{t-1}] + \theta_{t-1} \tag{50}$$

$$w_t - \tilde{w}_t = -\eta\sum_{\tau=0}^{t-1}\Delta_\tau \tag{51}$$

Let filtration $\mathfrak{F}_t = \sigma\{\xi_0, \cdots \xi_{t-1}\}$, and note $\sigma\{\Delta_0, \cdots, \Delta_t\} \subset \mathfrak{F}_t$, where $\sigma\{\cdot\}$ denotes the sigma field. Also, let event $\mathfrak{K}_t = \{\forall\tau \le t,\ \|\nabla\tilde{f}(\tilde{w}_\tau)\| \le \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta}),\ \|\tilde{w}_\tau - w_0\| \le \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta})\}$, and $\mathfrak{E}_t = \{\forall\tau \le t,\ \|\Delta_\tau\| \le \mu\eta\log^2\frac{1}{\eta}\}$, where μ is independent of (η, ζ), and will be specified later. Again, $\tilde{O}$ notation in this proof will never hide any dependence on μ. Clearly, we have $\mathfrak{K}_t \subset \mathfrak{K}_{t-1}$ ($\mathfrak{E}_t \subset \mathfrak{E}_{t-1}$), thus $1_{\mathfrak{K}_t} \le 1_{\mathfrak{K}_{t-1}}$ ($1_{\mathfrak{E}_t} \le 1_{\mathfrak{E}_{t-1}}$), where $1_{\mathfrak{K}}$ is the indicator function of event $\mathfrak{K}$.

We first need to carefully bounded all terms in Eq.(50), conditioned on event $\mathfrak{K}_{t-1} \cap \mathfrak{E}_{t-1}$, by Eq.(47), Eq.(48)), and Eq.(51), with probability 1, for all $t \le T \le O(\log d/\gamma_0\eta)$, we have:

$$\begin{aligned}\|(1 - \eta\mathcal{H})\Delta_{t-1}\| &\le \tilde{O}(\mu\eta\log^2\frac{1}{\eta}) \qquad \|\eta\mathcal{H}'_{t-1}(\Delta_{t-1} + \nabla\tilde{f}(\tilde{w}_{t-1}))\| \le \tilde{O}(\eta^2\log^2\frac{1}{\eta}) \\ \|\eta\mathcal{H}'_{t-1}\xi_{t-1}\| &\le \tilde{O}(\eta^{1.5}\log\frac{1}{\eta}) \qquad \|\theta_{t-1}\| \le \tilde{O}(\eta^2)\end{aligned} \tag{52}$$

Since event $\mathfrak{K}_{t-1} \subset \mathfrak{F}_{t-1}, \mathfrak{E}_{t-1} \subset \mathfrak{F}_{t-1}$ thus independent of ξ_{t-1}, we also have:

$$\begin{aligned}&\mathbb{E}[((1 - \eta\mathcal{H})\Delta_{t-1})^T\eta\mathcal{H}'_{t-1}\xi_{t-1}1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] \\ =&1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}((1 - \eta\mathcal{H})\Delta_{t-1})^T\eta\mathcal{H}'_{t-1}\mathbb{E}[\xi_{t-1} \mid \mathfrak{F}_{t-1}] = 0\end{aligned} \tag{53}$$

Therefore, from Eq.(50) and Eq.(52):

$$\begin{aligned}&\mathbb{E}[\|\Delta_t\|_2^2 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] \\ \le&\left[(1 + \eta\gamma_0)^2\|\Delta_{t-1}\|^2 + (1 + \eta\gamma_0)\|\Delta_{t-1}\|\tilde{O}(\eta^2\log^2\frac{1}{\eta}) + \tilde{O}(\eta^3\log^2\frac{1}{\eta})\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \\ \le&\left[(1 + \eta\gamma_0)^2\|\Delta_{t-1}\|^2 + \tilde{O}(\mu\eta^3\log^4\frac{1}{\eta})\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}\end{aligned} \tag{54}$$

Define

$$G_t = (1 + \eta\gamma_0)^{-2t}[\ \|\Delta_t\|^2 + \alpha\eta^2\log^4\frac{1}{\eta}\ ] \tag{55}$$

Then, when $\eta_{\max}$ is small enough, we have:

$$\mathbb{E}[G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] = (1+\eta\gamma_0)^{-2t}\left[\mathbb{E}[\|\Delta_t\|_2^2 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] + \alpha\eta^2\log^3\frac{1}{\eta}\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}$$
$$\leq (1+\eta\gamma_0)^{-2t}\left[(1+\eta\gamma_0)^2\|\Delta_{t-1}\|^2 + \tilde{O}(\mu\eta^3\log^4\frac{1}{\eta}) + \alpha\eta^2\log^4\frac{1}{\eta}\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}$$
$$\leq (1+\eta\gamma_0)^{-2t}\left[(1+\eta\gamma_0)^2\|\Delta_{t-1}\|^2 + (1+\eta\gamma_0)^2\alpha\eta^2\log^4\frac{1}{\eta}\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}$$
$$= G_{t-1}1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \leq G_{t-1}1_{\mathfrak{K}_{t-2}\cap\mathfrak{E}_{t-2}} \tag{56}$$

Therefore, we have $\mathbb{E}[G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] \leq G_{t-1}1_{\mathfrak{K}_{t-2}\cap\mathfrak{E}_{t-2}}$ which means $G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}$ is a supermartingale.

On the other hand, we have:

$$\Delta_t = (1-\eta H)\Delta_{t-1} - \eta\mathcal{H}'_{t-1}(\Delta_{t-1} + \nabla\tilde{f}(\tilde{w}_{t-1})) - \eta\mathcal{H}'_{t-1}\xi_{t-1} + \theta_{t-1} \tag{57}$$

Once conditional on filtration $\mathfrak{F}_{t-1}$, the first two terms are deterministic, and only the third and fourth term are random. Therefore, we know, with probability 1:

$$|\ \|\Delta_t\|_2^2 - \mathbb{E}[\|\Delta_t\|_2^2|\mathfrak{F}_{t-1}]\ |1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \leq \tilde{O}(\mu\eta^{2.5}\log^3\frac{1}{\eta}) \tag{58}$$

Where the main contribution comes from the product of the first term and third term. Then, with probability 1, we have:

$$|G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} - \mathbb{E}[G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}]|$$
$$= (1+2\eta\gamma_0)^{-2t}\cdot|\ \|\Delta_t\|_2^2 - \mathbb{E}[\|\Delta_t\|_2^2|\mathfrak{F}_{t-1}]\ |\cdot 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \leq \tilde{O}(\mu\eta^{2.5}\log^3\frac{1}{\eta}) = c_{t-1} \tag{59}$$

By Azuma-Hoeffding inequality, with probability less than $\tilde{O}(\eta^3)$, for $t \leq T \leq O(\log d/\gamma_0\eta)$:

$$G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} - G_0\cdot 1 > \tilde{O}(1)\sqrt{\sum_{\tau=0}^{t-1}c_\tau^2\log(\frac{1}{\eta})} = \tilde{O}(\mu\eta^2\log^4\frac{1}{\eta}) \tag{60}$$

This means there exist some $\tilde{C} = \tilde{O}(1)$ so that:

$$P\left(G_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \geq \tilde{C}\mu\eta^2\log^4\frac{1}{\eta}\right) \leq \tilde{O}(\eta^3) \tag{61}$$

By choosing $\mu > \tilde{C}$, this is equivalent to:

$$P\left(\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}\cap\left\{\|\Delta_t\|^2 \geq \mu^2\eta^2\log^4\frac{1}{\eta}\right\}\right) \leq \tilde{O}(\eta^3) \tag{62}$$

Therefore, combined with Lemma 19, we have:

$$P\left(\mathfrak{E}_{t-1}\cap\left\{\|\Delta_t\| \geq \mu\eta\log^2\frac{1}{\eta}\right\}\right)$$
$$= P\left(\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}\cap\left\{\|\Delta_t\| \geq \mu\eta\log^2\frac{1}{\eta}\right\}\right) + P\left(\overline{\mathfrak{K}}_{t-1}\cap\mathfrak{E}_{t-1}\cap\left\{\|\Delta_t\| \geq \mu\eta\log^2\frac{1}{\eta}\right\}\right)$$
$$\leq \tilde{O}(\eta^3) + P(\overline{\mathfrak{K}}_{t-1}) \leq \tilde{O}(\eta^3) \tag{63}$$

Finally, we know:

$$P(\overline{\mathfrak{E}}_t) = P\left(\mathfrak{E}_{t-1} \cap \left\{\|\Delta_t\| \geq \mu\eta \log^2\frac{1}{\eta}\right\}\right) + P(\overline{\mathfrak{E}}_{t-1}) \leq \tilde{O}(\eta^3) + P(\overline{\mathfrak{E}}_{t-1}) \tag{64}$$

Because $P(\overline{\mathfrak{E}}_0) = 0$, and $T \leq \tilde{O}(\frac{1}{\eta})$, we have $P(\overline{\mathfrak{E}}_T) \leq \tilde{O}(\eta^2)$. Due to Eq.(51), we have $\|w_t - \tilde{w}_t\| \leq \eta\sum_{\tau=0}^{t-1}\|\Delta_\tau\|$, then by the definition of $\mathfrak{E}_T$, we finish the proof.

■

Using the two lemmas above we are ready to prove Lemma 17

Proof [Proof of Lemma 17] Let $\tilde{f}$ and $\{\tilde{w}_t\}$ be defined as in Lemma 19. and also let $\lambda_{\min}(\mathcal{H}(w_0)) = -\gamma_0$. Since $\mathcal{H}(w)$ is ρ-Lipschitz, for any w, w_0, we have:

$$f(w) \leq f(w_0) + \nabla f(w_0)^T(w - w_0) + \frac{1}{2}(w-w_0)^T\mathcal{H}(w_0)(w-w_0) + \frac{\rho}{6}\|w - w_0\|^3 \tag{65}$$

Denote $\tilde{\delta} = \tilde{w}_T - w_0$ and $\delta = w_T - \tilde{w}_T$, we have:

$$\begin{aligned}
f(w_T) - f(w_0) &\leq \left[\nabla f(w_0)^T(w_T - w_0) + \frac{1}{2}(w_T-w_0)^T\mathcal{H}(w_0)(w_T-w_0) + \frac{\rho}{6}\|w_T - w_0\|^3\right]\\
&= \left[\nabla f(w_0)^T(\tilde{\delta}+\delta) + \frac{1}{2}(\tilde{\delta}+\delta)^T\mathcal{H}(\tilde{\delta}+\delta) + \frac{\rho}{6}\|\tilde{\delta}+\delta\|^3\right]\\
&= \left[\nabla f(w_0)^T\tilde{\delta} + \frac{1}{2}\tilde{\delta}^T\mathcal{H}\tilde{\delta}\right] + \left[\nabla f(w_0)^T\delta + \tilde{\delta}^T\mathcal{H}\delta + \frac{1}{2}\delta^T\mathcal{H}\delta + \frac{\rho}{6}\|\tilde{\delta}+\delta\|^3\right]
\end{aligned} \tag{66}$$

Where $\mathcal{H} = \mathcal{H}(w_0)$. Denote $\tilde{\Lambda} = \nabla f(w_0)^T\tilde{\delta} + \frac{1}{2}\tilde{\delta}^T\mathcal{H}\tilde{\delta}$ be the first term, and $\Lambda = \nabla f(w_0)^T\delta + \tilde{\delta}^T\mathcal{H}\delta + \frac{1}{2}\delta^T\mathcal{H}\delta + \frac{\rho}{6}\|\tilde{\delta}+\delta\|^3$ be the second term. We have $f(w_T) - f(w_0) \leq \tilde{\Lambda} + \Lambda$.

Let $\mathfrak{E}_t = \{\forall\tau \leq t, \|\tilde{w}_\tau - w_0\| \leq \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta}),\ \|w_t - \tilde{w}_t\| \leq \tilde{O}(\eta\log^2\frac{1}{\eta})\}$, by the result of Lemma 19 and Lemma 20, we know $P(\mathfrak{E}_T) \geq 1 - \tilde{O}(\eta^2)$. Then, clearly, we have:

$$\begin{aligned}
\mathbb{E}f(w_T) - f(w_0) =& \mathbb{E}[f(w_T) - f(w_0)]1_{\mathfrak{E}_T} + \mathbb{E}[f(w_T) - f(w_0)]1_{\overline{\mathfrak{E}}_T}\\
\leq& \mathbb{E}\tilde{\Lambda}1_{\mathfrak{E}_T} + \mathbb{E}\Lambda 1_{\mathfrak{E}_T} + \mathbb{E}[f(w_T) - f(w_0)]1_{\overline{\mathfrak{E}}_T}\\
=& \mathbb{E}\tilde{\Lambda} + \mathbb{E}\Lambda 1_{\mathfrak{E}_T} + \mathbb{E}[f(w_T) - f(w_0)]1_{\overline{\mathfrak{E}}_T} - \mathbb{E}\tilde{\Lambda}1_{\overline{\mathfrak{E}}_T}
\end{aligned} \tag{67}$$

We will carefully caculate $\mathbb{E}\tilde{\Lambda}$ term first, and then bound remaining term as "perturbation" to first term.

Let $\lambda_1, \cdots, \lambda_d$ be the eigenvalues of $\mathcal{H}$. By the result of lemma 19 and simple linear algebra, we have:

$$\begin{aligned}
\mathbb{E}\tilde{\Lambda} &= -\frac{\eta}{2}\sum_{i=1}^{d}\sum_{\tau=0}^{2T-1}(1-\eta\lambda_i)^\tau|\nabla_i f(w_0)|^2 + \frac{1}{2}\sum_{i=1}^{d}\lambda_i\sum_{\tau=0}^{T-1}(1-\eta\lambda_i)^{2\tau}\eta^2\sigma^2\\
&\leq \frac{1}{2}\sum_{i=1}^{d}\lambda_i\sum_{\tau=0}^{T-1}(1-\eta\lambda_i)^{2\tau}\eta^2\sigma^2\\
&\leq \frac{\eta^2\sigma^2}{2}\left[\frac{d-1}{\eta} - \gamma_0\sum_{\tau=0}^{T-1}(1+\eta\gamma_0)^{2\tau}\right] \leq -\frac{\eta\sigma^2}{2}
\end{aligned} \tag{68}$$

The last inequality is directly implied by the choice of T as in Eq.(35). Also, by Eq.(35), we also immediately have that $T = O(\log d/\gamma_0\eta) \leq O(\log d/\gamma\eta)$. Therefore, by choose $T_{max} = O(\log d/\gamma\eta)$ with large enough constant, we have $T \leq T_{max} = O(\log d/\gamma\eta)$.

For bounding the second term, by definition of $\mathfrak{E}_t$, we have:

$$\mathbb{E}\Lambda 1_{\mathfrak{E}_T} = \mathbb{E}\left[\nabla f(w_0)^T\delta + \tilde{\delta}^T\mathcal{H}\delta + \frac{1}{2}\delta^T\mathcal{H}\delta + \frac{\rho}{6}\|\tilde{\delta}+\delta\|^3\right]1_{\mathfrak{E}_T} \leq \tilde{O}(\eta^{1.5}\log^3\frac{1}{\eta}) \tag{69}$$

On the other hand, since noise is bounded as $\|\xi\| \leq \tilde{O}(1)$, from the results of Lemma 19, it's easy to show $\|\tilde{w} - w_0\| = \|\tilde{\delta}\| \leq \tilde{O}(1)$ is also bounded with probability 1. Recall the assumption that function f is also bounded, then we have:

$$\begin{aligned}&\mathbb{E}[f(w_T) - f(w_0)]1_{\overline{\mathfrak{E}}_T} - \mathbb{E}\tilde{\Lambda}1_{\overline{\mathfrak{E}}_T}\\ =&\mathbb{E}[f(w_T) - f(w_0)]1_{\overline{\mathfrak{E}}_T} - \mathbb{E}\left[\nabla f(w_0)^T\tilde{\delta} + \frac{1}{2}\tilde{\delta}^T\mathcal{H}\tilde{\delta}\right]1_{\overline{\mathfrak{E}}_T} \leq \tilde{O}(1)P(\overline{\mathfrak{E}}_T) \leq \tilde{O}(\eta^2)\end{aligned} \tag{70}$$

Finally, substitute Eq.(68), Eq.(69) and Eq.(70) into Eq.(67), we finish the proof. ∎

Finally, we combine three cases to prove the main theorem.

Proof [Proof of Theorem 14] Let's set $\mathcal{L}_1 = \{w \mid \|\nabla f(w)\| \geq \sqrt{2\eta\sigma^2\beta d}\}$, $\mathcal{L}_2 = \{w \mid \|\nabla f(w)\| \leq \sqrt{2\eta\sigma^2\beta d}$ and $\lambda_{\min}(\mathcal{H}(w)) \leq -\gamma\}$, and $\mathcal{L}_3 = \mathcal{L}_1^c \cup \mathcal{L}_2^c$. By choosing small enough $\eta_{\max}$, we could make $\sqrt{2\eta\sigma^2\beta d} < \min\{\epsilon, \alpha\delta\}$. Under this choice, we know from Definition 5 of $(\alpha,\gamma,\epsilon,\delta)$-strict saddlethat $\mathcal{L}_3$ is the locally α-strongly convex region which is $\tilde{O}(\sqrt{\eta})$-close to some local minimum.

We shall first prove that within $\tilde{O}(\frac{1}{\eta^2}\log\frac{1}{\zeta})$ steps with probability at least $1-\zeta/2$ one of w_t is in $\mathcal{L}_3$. Then by Lemma 16 we know with probability at most $\zeta/2$ there exists a w_t that is in $\mathcal{L}_3$ but the last point is not. By union bound we will get the main result.

To prove within $\tilde{O}(\frac{1}{\eta^2}\log\frac{1}{\zeta})$ steps with probability at least $1-\zeta/2$ one of w_t is in $\mathcal{L}_3$, we first show starting from any point, in $\tilde{O}(\frac{1}{\eta^2})$ steps with probability at least $1/2$ one of w_t is in $\mathcal{L}_3$. Then we can repeat this $\log 1/\zeta$ times to get the high probability result.

Define stochastic process $\{\tau_i\}$ s.t. $\tau_0 = 0$, and

$$\tau_{i+1} = \begin{cases}\tau_i + 1 & \text{if } w_{\tau_i} \in \mathcal{L}_1 \cup \mathcal{L}_3\\ \tau_i + T(w_{\tau_i}) & \text{if } w_{\tau_i} \in \mathcal{L}_2\end{cases} \tag{71}$$

Where $T(w_{\tau_i})$ is defined by Eq.(35) with $\gamma_0 = \lambda_{\min}(\mathcal{H}(w_{\tau_i}))$and we know $T \leq T_{max} = \tilde{O}(\frac{1}{\eta})$.

By Lemma 15 and Lemma 17, we know:

$$\mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \in \mathcal{L}_1, \mathfrak{F}_{\tau_i - 1}] = \mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \in \mathcal{L}_1] \leq -\tilde{O}(\eta^2) \tag{72}$$

$$\mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \in \mathcal{L}_2, \mathfrak{F}_{\tau_i - 1}] = \mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \in \mathcal{L}_2] \leq -\tilde{O}(\eta) \tag{73}$$

Therefore, combine above equation, we have:

$$\mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \notin \mathcal{L}_3, \mathfrak{F}_{\tau_i - 1}] = \mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|w_{\tau_i} \notin \mathcal{L}_3] \leq -(\tau_{i+1} - \tau_i)\tilde{O}(\eta^2) \tag{74}$$

Define event $\mathfrak{E}_i = \{\exists j \leq i,\ w_{\tau_j} \in \mathcal{L}_3\}$, clearly $\mathfrak{E}_i \subset \mathfrak{E}_{i+1}$, thus $P(\mathfrak{E}_i) \leq P(\mathfrak{E}_{i+1})$. Finally, consider $f(w_{\tau_{i+1}})1_{\mathfrak{E}_i}$, we have:

$$\begin{aligned}\mathbb{E}f(w_{\tau_{i+1}})1_{\mathfrak{E}_i} - \mathbb{E}f(w_{\tau_i})1_{\mathfrak{E}_{i-1}} &\leq B\cdot P(\mathfrak{E}_i - \mathfrak{E}_{i-1}) + \mathbb{E}[f(w_{\tau_{i+1}}) - f(w_{\tau_i})|\overline{\mathfrak{E}_i}]\cdot P(\overline{\mathfrak{E}_i})\\ &\leq B\cdot P(\mathfrak{E}_i - \mathfrak{E}_{i-1}) - (\tau_{i+1} - \tau_i)\tilde{O}(\eta^2)P(\overline{\mathfrak{E}_i})\end{aligned} \tag{75}$$

Therefore, by summing up over i, we have:

$$\mathbb{E}f(w_{\tau_i})1_{\mathfrak{E}_i} - f(w_0) \le BP(\mathfrak{E}_i) - \tau_i\tilde{O}(\eta^2)P(\overline{\mathfrak{E}_i}) \le B - \tau_i\tilde{O}(\eta^2)P(\overline{\mathfrak{E}_i}) \tag{76}$$

Since $|f(w_{\tau_i})1_{\mathfrak{E}_i}| < B$ is bounded, as τ_i grows to as large as $\frac{6B}{\eta^2}$, we must have $P(\overline{\mathfrak{E}_i}) < \frac{1}{2}$. That is, after $\tilde{O}(\frac{1}{\eta^2})$ steps, with at least probability $1/2$, $\{w_t\}$ have at least enter $\mathcal{L}_3$ once. Since this argument holds for any starting point, we can repeat this $\log 1/\zeta$ times and we know after $\tilde{O}(\frac{1}{\eta^2}\log 1/\zeta)$ steps, with probability at least $1-\zeta/2$, $\{w_t\}$ have at least enter $\mathcal{L}_3$ once.

Combining with Lemma 16, and by union bound we know after $\tilde{O}(\frac{1}{\eta^2}\log 1/\zeta)$ steps, with probability at least $1-\zeta$, w_t will be in the $\tilde{O}(\sqrt{\eta\log\frac{1}{\eta\zeta}})$ neigborhood of some local minimum. ■

Appendix B. Detailed Analysis for Section 3 in Constrained Case

So far, we have been discussed all about unconstrained problem. In this section we extend our result to equality constraint problems under some mild conditions.

Consider the equality constrained optimization problem:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & c_i(w) = 0, \qquad i = 1, \cdots, m \end{aligned} \tag{77}$$

Define the feasible set as the set of points that satisfy all the constraints $\mathcal{W} = \{w \mid c_i(w) = 0;\ i = 1, \cdots, m\}$.

In this case, the algorithm we are running is Projected Noisy Gradient Descent. Let function $\Pi_{\mathcal{W}}(v)$ to be the projection to the feasible set where the projection is defined as the global solution of $\min_{w\in\mathcal{W}}\|v-w\|^2$.

With same argument as in the unconstrained case, we could slightly simplify and convert it to standard projected stochastic gradient descent (PSGD) with update equation:

$$v_t = w_{t-1} - \eta\nabla f(w_{t-1}) + \xi_{t-1} \tag{78}$$

$$w_t = \Pi_{\mathcal{W}}(v_t) \tag{79}$$

As in unconstrained case, we are interested in noise ξ is i.i.d satisfying $\mathbb{E}\xi = 0$, $\mathbb{E}\xi\xi^T = \sigma^2 I$ and $\|\xi\| \le Q$ almost surely. Our proof can be easily extended to Algorithm 2 with $\frac{1}{d}I \preceq \mathbb{E}\xi\xi^T \preceq (Q+\frac{1}{d})I$. In this section we first introduce basic tools for handling constrained optimization problems (most these materials can be found in Wright and Nocedal (1999)), then we prove some technical lemmas that are useful for dealing with the projection step in PSGD, finally we point out how to modify the previous analysis.

B.1. Preliminaries

Often for constrained optimization problems we want the constraints to satisfy some regularity conditions. LICQ (linear independent constraint quantification) is a common assumption in this context.

Definition 21 (LICQ) *In equality-constraint problem Eq.(77), given a point w, we say that the linear independence constraint qualification (LICQ) holds if the set of constraint gradients $\{\nabla c_i(x), i = 1, \cdots, m\}$ is linearly independent.*

In constrained optimization, we can locally transform it to an unconstrained problem by introducing Lagrangian multipliers. The Langrangian $\mathcal{L}$ can be written as

$$\mathcal{L}(w,\lambda) = f(w) - \sum_{i=1}^{m} \lambda_i c_i(w) \tag{80}$$

Then, if LICQ holds for all $w \in \mathcal{W}$, we can properly define function $\lambda^*(\cdot)$ to be:

$$\lambda^*(w) = \arg\min_{\lambda} \|\nabla f(w) - \sum_{i=1}^{m} \lambda_i \nabla c_i(w)\| = \arg\min_{\lambda} \|\nabla_w \mathcal{L}(w,\lambda)\| \tag{81}$$

where $\lambda^*(\cdot)$ can be calculated analytically: let matrix $C(w) = (\nabla c_1(w), \cdots, \nabla c_m(w))$, then we have:

$$\lambda^*(w) = C(w)^\dagger \nabla f(w) = (C(w)^T C(w))^{-1} C(w)^T \nabla f(w) \tag{82}$$

where $(\cdot)^\dagger$ is Moore-Penrose pseudo-inverse.

In our setting we need a stronger regularity condition which we call robust LICQ (RLICQ).

Definition 22 (α_c-RLICQ) *In equality-constraint problem Eq.(77), given a point w, we say that α_c-robust linear independence constraint qualification (α_c-RLICQ) holds if the minimum singular value of matrix $C(w) = (\nabla c_1(w), \cdots, \nabla c_m(w))$ is greater or equal to α_c, that is $\sigma_{\min}(C(w)) \geq \alpha_c$.*

Remark 23 *Given a point $w \in \mathcal{W}$, α_c-RLICQ implies LICQ. While LICQ holds for all $w \in \mathcal{W}$ is a necessary condition for $\lambda^*(w)$ to be well-defined; it's easy to check that α_c-RLICQ holds for all $w \in \mathcal{W}$ is a necessary condition for $\lambda^*(w)$ to be bounded. Later, we will also see α_c-RLICQ combined with the smoothness of $\{c_i(w)\}_{i=1}^m$ guarantee the curvature of constraint manifold to be bounded everywhere.*

Note that we require this condition in order to provide a quantitative bound, without this assumption there can be cases that are exponentially close to a function that does not satisfy LICQ.

We can also write down the first-order and second-order partial derivative of Lagrangian $\mathcal{L}$ at point $(w, \lambda^*(w))$:

$$\chi(w) = \nabla_w \mathcal{L}(w,\lambda)|_{(w,\lambda^*(w))} = \nabla f(w) - \sum_{i=1}^{m} \lambda_i^*(w) \nabla c_i(w) \tag{83}$$

$$\mathfrak{M}(w) = \nabla^2_{ww} \mathcal{L}(w,\lambda)|_{(w,\lambda^*(w))} = \nabla^2 f(w) - \sum_{i=1}^{m} \lambda_i^*(w) \nabla^2 c_i(w) \tag{84}$$

Definition 24 (Tangent Space and Normal Space) *Given a feasible point $w \in \mathcal{W}$, define its corresponding Tangent Space to be $\mathcal{T}(w) = \{v \mid \nabla c_i(w)^T v = 0;\ i = 1, \cdots, m\}$, and Normal Space to be $\mathcal{T}^c(w) = span\{\nabla c_1(w) \cdots, \nabla c_m(w)\}$*

If $w \in \mathcal{R}^d$, and we have m constraint satisfying α_c-RLICQ , the tangent space would be a linear subspace with dimension $d - m$; and the normal space would be a linear subspace with dimension m. We also know immediately that $\chi(w)$ defined in Eq.(83) has another interpretation: it's the component of gradient $\nabla f(w)$ in tangent space.

Also, it's easy to see the normal space $\mathcal{T}^c(w)$ is the orthogonal complement of $\mathcal{T}$. We can also define the projection matrix of any vector onto tangent space (or normal space) to be $P_{\mathcal{T}(w)}$ (or $P_{\mathcal{T}^c(w)}$). Then, clearly, both $P_{\mathcal{T}(w)}$ and $P_{\mathcal{T}^c(w)}$ are orthoprojector, thus symmetric. Also by Pythagorean theorem, we have:

$$\|v\|^2 = \|P_{\mathcal{T}(w)} v\|^2 + \|P_{\mathcal{T}^c(w)} v\|^2, \qquad \forall v \in \mathbb{R}^d \tag{85}$$

Taylor Expansion Let $w, w_0 \in \mathcal{W}$, and fix $\lambda^* = \lambda^*(w_0)$ independent of w, assume $\nabla^2_{ww}\mathcal{L}(w, \lambda^*)$ is ρ_L-Lipschitz, that is $\|\nabla^2_{ww}\mathcal{L}(w_1, \lambda^*) - \nabla^2_{ww}\mathcal{L}(w_2, \lambda^*)\| \leq \rho_L\|w_1 - w_2\|$ By Taylor expansion, we have:

$$\begin{aligned}\mathcal{L}(w, \lambda^*) \leq &\mathcal{L}(w_0, \lambda^*) + \nabla_w\mathcal{L}(w_0, \lambda^*)^T(w - w_0) \\ &+ \frac{1}{2}(w - w_0)^T\nabla^2_{ww}\mathcal{L}(w_0, \lambda^*)(w - w_0) + \frac{\rho_L}{6}\|w - w_0\|^3\end{aligned} \tag{86}$$

Since w, w_0 are feasible, we know: $\mathcal{L}(w, \lambda^*) = f(w)$ and $\mathcal{L}(w_0, \lambda^*) = f(w_0)$, this gives:

$$f(w) \leq f(w_0) + \chi(w_0)^T(w - w_0) + \frac{1}{2}(w - w_0)^T\mathfrak{M}(w_0)(w - w_0) + \frac{\rho_L}{6}\|w - w_0\|^3 \tag{87}$$

Derivative of $\chi(w)$ By taking derative of $\chi(w)$ again, we know the change of this tangent gradient can be characterized by:

$$\nabla\chi(w) = \mathcal{H} - \sum_{i=1}^{m}\lambda_i^*(w)\nabla^2 c_i(w) - \sum_{i=1}^{m}\nabla c_i(w)\nabla\lambda_i^*(w)^T \tag{88}$$

Denote

$$\mathfrak{N}(w) = -\sum_{i=1}^{m}\nabla c_i(w)\nabla\lambda_i^*(w)^T \tag{89}$$

We immediately know that $\nabla\chi(w) = \mathfrak{M}(w) + \mathfrak{N}(w)$.

Remark 25 *The additional term $\mathfrak{N}(w)$ is not necessary to be even symmetric in general. This is due to the fact that $\chi(w)$ may not be the gradient of any scalar function. However, $\mathfrak{N}(w)$ has an important property that is: for any vector $v \in \mathbb{R}^d$, $\mathfrak{N}(w)v \in \mathcal{T}^c(w)$.*

Finally, for completeness, we state here the first/second-order necessary (or sufficient) conditions for optimality. Please refer to Wright and Nocedal (1999) for the proof of those theorems.

Theorem 26 (First-Order Necessary Conditions) *In equality constraint problem Eq.(77), suppose that $w^\dagger$ is a local solution, and that the functions f and c_i are continuously differentiable, and that the LICQ holds at $w^\dagger$. Then there is a Lagrange multiplier vector $\lambda^\dagger$, such that:*

$$\nabla_w\mathcal{L}(w^\dagger, \lambda^\dagger) = 0 \tag{90}$$

$$c_i(w^\dagger) = 0, \qquad \text{for } i = 1, \cdots, m \tag{91}$$

These conditions are also usually referred as Karush-Kuhn-Tucker (KKT) conditions.

Theorem 27 (Second-Order Necessary Conditions) *In equality constraint problem Eq.(77), suppose that $w^\dagger$ is a local solution, and that the LICQ holds at $w^\dagger$. Let $\lambda^\dagger$ Lagrange multiplier vector for which the KKT conditions are satisfied. Then:*

$$v^T\nabla^2_{xx}\mathcal{L}(w^\dagger, \lambda^\dagger)v \geq 0 \qquad \text{for all } v \in \mathcal{T}(w^\dagger) \tag{92}$$

Theorem 28 (Second-Order Sufficient Conditions) *In equality constraint problem Eq.(77), suppose that for some feasible point $w^\dagger \in \mathbb{R}^d$, and there's Lagrange multiplier vector $\lambda^\dagger$ for which the KKT conditions are satisfied. Suppose also that:*

$$v^T\nabla^2_{xx}\mathcal{L}(w^\dagger, \lambda^\dagger)v > 0 \qquad \text{for all } v \in \mathcal{T}(w^\dagger), v \neq 0 \tag{93}$$

Then $w^\dagger$ is a strict local solution.

Remark 29 *By definition Eq.(82), we know immediately $\lambda^*(w^\dagger)$ is one of valid Lagrange multipliers $\lambda^\dagger$ for which the KKT conditions are satisfied. This means $\chi(w^\dagger) = \nabla_w \mathcal{L}(w^\dagger, \lambda^\dagger)$ and $\mathfrak{M}(w^\dagger) = \mathcal{L}(w^\dagger, \lambda^\dagger)$.*

Therefore, Theorem 26, 27, 28 gives strong implication that $\chi(w)$ and $\mathfrak{M}(w)$ are the right thing to look at, which are in some sense equivalent to $\nabla f(w)$ and $\nabla^2 f(w)$ in unconstrained case.

B.2. Geometrical Lemmas Regarding Constraint Manifold

Since in equality constraint problem, at each step of PSGD, we are effectively considering the local manifold around feasible point w_{t-1}. In this section, we provide some technical lemmas relating to the geometry of constraint manifold in preparsion for the proof of main theorem in equality constraint case.

We first show if two points are close, then the projection in the normal space is much smaller than the projection in the tangent space.

Lemma 30 *Suppose the constraints $\{c_i\}_{i=1}^m$ are β_i-smooth, and α_c-RLICQ holds for all $w \in \mathcal{W}$. Then, let $\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2} = \frac{1}{R^2}$, for any $w, w_0 \in \mathcal{W}$, let $\mathcal{T}_0 = \mathcal{T}(w_0)$, then*

$$\|P_{\mathcal{T}_0^c}(w - w_0)\| \le \frac{1}{2R}\|w - w_0\|^2 \tag{94}$$

Furthermore, if $\|w - w_0\| < R$ holds, we additionally have:

$$\|P_{\mathcal{T}_0^c}(w - w_0)\| \le \frac{\|P_{\mathcal{T}_0}(w - w_0)\|^2}{R} \tag{95}$$

Proof First, since for any vector $\hat{v} \in \mathcal{T}_0$, we have $\|C(w_0)^T \hat{v}\| = 0$, then by simple linear algebra, it's easy to show:

$$\begin{aligned}\|C(w_0)^T(w - w_0)\|^2 =& \|C(w_0)^T P_{\mathcal{T}_0^c}(w - w_0)\|^2 \ge \sigma_{\min}^2 \|P_{\mathcal{T}_0^c}(w - w_0)\|^2 \\ \ge& \alpha_c^2 \|P_{\mathcal{T}_0^c}(w - w_0)\|^2\end{aligned} \tag{96}$$

On the other hand, by β_i-smooth, we have:

$$|c_i(w) - c_i(w_0) - \nabla c_i(w_0)^T(w - w_0)| \le \frac{\beta_i}{2}\|w - w_0\|^2 \tag{97}$$

Since w, w_0 are feasible points, we have $c_i(w) = c_i(w_0) = 0$, which gives:

$$\|C(w_0)^T(w - w_0)\|^2 = \sum_{i=1}^m (\nabla c_i(w_0)^T(w - w_0))^2 \le \sum_{i=1}^m \frac{\beta_i^2}{4}\|w - w_0\|^4 \tag{98}$$

Combining Eq.(96) and Eq.(98), and the definition of R, we have:

$$\|P_{\mathcal{T}_0^c}(w - w_0)\|^2 \le \frac{1}{4R^2}\|w - w_0\|^4 = \frac{1}{4R^2}(\|P_{\mathcal{T}_0^c}(w - w_0)\|^2 + \|P_{\mathcal{T}_0}(w - w_0)\|^2)^2 \tag{99}$$

Solving this second-order inequality gives two solution

$$\|P_{\mathcal{T}_0^c}(w - w_0)\| \le \frac{\|P_{\mathcal{T}_0}(w - w_0)\|^2}{R} \quad \text{or} \quad \|P_{\mathcal{T}_0^c}(w - w_0)\| \ge R \tag{100}$$

By assumption, we know $\|w - w_0\| < R$ (so the second case cannot be true), which finishes the proof. ■

Here, we see the $\sqrt{\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2}} = \frac{1}{R}$ serves as a upper bound of the curvatures on the constraint manifold, and equivalently, R serves as a lower bound of the radius of curvature. α_c-RLICQ and smoothness guarantee that the curvature is bounded.

Next we show the normal/tangent space of nearby points are close.

Lemma 31 *Suppose the constraints $\{c_i\}_{i=1}^m$ are β_i-smooth, and α_c-RLICQ holds for all $w \in \mathcal{W}$. Let $\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2} = \frac{1}{R^2}$, for any $w, w_0 \in \mathcal{W}$, let $\mathcal{T}_0 = \mathcal{T}(w_0)$. Then for all $\hat{v} \in \mathcal{T}(w)$ so that $\|\hat{v}\| = 1$, we have*

$$\|P_{\mathcal{T}_0^c} \cdot \hat{v}\| \leq \frac{\|w - w_0\|}{R} \tag{101}$$

Proof With similar calculation as Eq.(96), we immediately have:

$$\|P_{\mathcal{T}_0^c} \cdot \hat{v}\|^2 \leq \frac{\|C(w_0)^T \hat{v}\|^2}{\sigma_{\min}^2(C(w))} \leq \frac{\|C(w_0)^T \hat{v}\|^2}{\alpha_c^2} \tag{102}$$

Since $\hat{v} \in \mathcal{T}(w)$, we have $C(w)^T \hat{v} = 0$, combined with the fact that $\hat{v}$ is a unit vector, we have:

$$\begin{aligned}\|C(w_0)^T \hat{v}\|^2 =& \|[C(w_0) - C(w)]^T \hat{v}\|^2 = \sum_{i=1}^m ([\nabla c_i(w_0) - \nabla c_i(w)]^T \hat{v})^2 \\ \leq& \sum_{i=1}^m \|\nabla c_i(w_0) - \nabla c_i(w)\|^2 \|\hat{v}\|^2 \leq \sum_{i=1}^m \beta_i^2 \|w_0 - w\|^2\end{aligned} \tag{103}$$

Combining Eq.(102) and Eq.(103), and the definition of R, we concludes the proof. ■

Lemma 32 *Suppose the constraints $\{c_i\}_{i=1}^m$ are β_i-smooth, and α_c-RLICQ holds for all $w \in \mathcal{W}$. Let $\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2} = \frac{1}{R^2}$, for any $w, w_0 \in \mathcal{W}$, let $\mathcal{T}_0 = \mathcal{T}(w_0)$. Then for all $\hat{v} \in \mathcal{T}^c(w)$ so that $\|\hat{v}\| = 1$, we have*

$$\|P_{\mathcal{T}_0} \cdot \hat{v}\| \leq \frac{\|w - w_0\|}{R} \tag{104}$$

Proof By definition of projection, clearly, we have $P_{\mathcal{T}_0} \cdot \hat{v} + P_{\mathcal{T}_0^c} \cdot \hat{v} = \hat{v}$. Since $\hat{v} \in \mathcal{T}^c(w)$, without loss of generality, assume $\hat{v} = \sum_{i=1}^m \lambda_i \nabla c_i(w)$. Define $\tilde{d} = \sum_{i=1}^m \lambda_i \nabla c_i(w_0)$, clearly $\tilde{d} \in \mathcal{T}_0^c$. Since projection gives the closest point in subspace, we have:

$$\begin{aligned}\|P_{\mathcal{T}_0} \cdot \hat{v}\| =& \|P_{\mathcal{T}_0^c} \cdot \hat{v} - \hat{v}\| \leq \|\tilde{d} - \hat{v}\| \\ \leq& \sum_{i=1}^m \lambda_i \|\nabla c_i(w_0) - \nabla c_i(w)\| \leq \sum_{i=1}^m \lambda_i \beta_i \|w_0 - w\|\end{aligned} \tag{105}$$

On the other hand, let $\lambda = (\lambda_1, \cdots, \lambda_m)^T$, we know $C(w)\lambda = \hat{v}$, thus:

$$\lambda = C(w)^\dagger \hat{v} = (C(w)^T C(w))^{-1} C(w)^T \hat{v} \tag{106}$$

Therefore, by α_c-RLICQ and the fact $\hat{v}$ is unit vector, we know: $\|\lambda\| \leq \frac{1}{\alpha_c}$. Combined with Eq.(105), we finished the proof. ∎

Using the previous lemmas, we can then prove that: starting from any point w_0 on constraint manifold, the result of adding any small vector v and then projected back to feasible set, is not very different from the result of adding $P_{\mathcal{T}(w_0)}v$.

Lemma 33 *Suppose the constraints $\{c_i\}_{i=1}^m$ are β_i-smooth, and α_c-RLICQ holds for all $w \in \mathcal{W}$. Let $\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2} = \frac{1}{R^2}$, for any $w_0 \in \mathcal{W}$, let $\mathcal{T}_0 = \mathcal{T}(w_0)$. Then let $w_1 = w_0 + \eta\hat{v}$, and $w_2 = w_0 + \eta P_{\mathcal{T}_0} \cdot \hat{v}$, where $\hat{v} \in \mathbb{S}^{d-1}$ is a unit vector. Then, we have:*

$$\|\Pi_{\mathcal{W}}(w_1) - w_2\| \leq \frac{4\eta^2}{R} \tag{107}$$

Where projection $\Pi_{\mathcal{W}}(w)$ is defined as the closet point to w on feasible set $\mathcal{W}$.

Proof First, note that $\|w_1 - w_0\| = \eta$, and by definition of projection, there must exist a project $\Pi_{\mathcal{W}}(w)$ inside the ball $\mathbb{B}_\eta(w_1) = \{w \mid \|w - w_1\| \leq \eta\}$.

Denote $u_1 = \Pi_{\mathcal{W}}(w_1)$, and clearly $u_1 \in \mathcal{W}$. we can formulate u_1 as the solution to following constrained optimization problems:

$$\begin{aligned} \min_u \quad & \|w_1 - u\|^2 \\ \text{s.t.} \quad & c_i(u) = 0, \qquad i = 1, \cdots, m \end{aligned} \tag{108}$$

Since function $f(u) = \|w_1 - u\|^2$ and $c_i(u)$ are continuously differentiable by assumption, and the condition α_c-RLICQ holds for all $w \in \mathcal{W}$ implies that LICQ holds for u_1. Therefore, by Karush-Kuhn-Tucker necessary conditions, we immediately know $(w_1 - u_1) \in \mathcal{T}^c(u_1)$.

Since $u_1 \in \mathbb{B}_\eta(w_1)$, we know $\|w_0 - u_1\| \leq 2\eta$, by Lemma 32, we immediately have:

$$\|P_{\mathcal{T}_0}(w_1 - u_1)\| = \frac{\|P_{\mathcal{T}_0}(w_1 - u_1)\|}{\|w_1 - u_1\|}\|w_1 - u_1\| \leq \frac{1}{R}\|w_0 - u_1\| \cdot \|w_1 - u_1\| \leq \frac{2}{R}\eta^2 \tag{109}$$

Let $v_1 = w_0 + P_{\mathcal{T}_0}(u_1 - w_0)$, we have:

$$\begin{aligned} \|v_1 - w_2\| =& \|(v_1 - w_0) - (w_2 - w_0)\| = \|P_{\mathcal{T}_0}(u_1 - w_0) - P_{\mathcal{T}_0}(w_1 - w_0)\| \\ =& \|P_{\mathcal{T}_0}(w_1 - u_1)\| \leq \frac{2}{R}\eta^2 \end{aligned} \tag{110}$$

On the other hand by Lemma 30, we have:

$$\|u_1 - v_1\| = \|P_{\mathcal{T}_0^c}(u_1 - w_0)\| \leq \frac{1}{2R}\|u_1 - w_0\|^2 \leq \frac{2}{R}\eta^2 \tag{111}$$

Combining Eq.(110) and Eq.(111), we finished the proof. ∎

B.3. Main Theorem

Now we are ready to prove the main theorems. First we revise the definition of strict saddle in the constrained case.

Definition 34 *A twice differentiable function $f(w)$ with constraints $c_i(w)$ is $(\alpha, \gamma, \epsilon, \delta)$-*strict saddle*, if for any point w one of the following is true*

1. $\|\chi(w)\| \geq \epsilon$.
2. $\hat{v}^T \mathfrak{M}(w)\hat{v} \leq -\gamma$ *for some* $\hat{v} \in \mathcal{T}(w)$, $\|\hat{v}\| = 1$
3. *There is a local minimum $w^\star$ such that $\|w - w^\star\| \leq \delta$, and for all w' in the 2δ neighborhood of $w^\star$, we have $\hat{v}^T \mathfrak{M}(w')\hat{v} \geq \alpha$ for all $\hat{v} \in \mathcal{T}(w')$, $\|\hat{v}\| = 1$*

Next, we prove a equivalent formulation for PSGD.

Lemma 35 *Suppose the constraints $\{c_i\}_{i=1}^m$ are β_i-smooth, and α_c-RLICQ holds for all $w \in \mathcal{W}$. Furthermore, if function f is L-Lipschitz, and the noise ξ is bounded, then running PSGD as in Eq.(78) is equivalent to running:*

$$w_t = w_{t-1} - \eta \cdot (\chi(w_{t-1}) + P_{\mathcal{T}(w_{t-1})}\xi_{t-1}) + \iota_{t-1} \tag{112}$$

where ι is the correction for projection, and $\|\iota\| \leq \tilde{O}(\eta^2)$.

Proof Lemma 35 is a direct corollary of Lemma 33. ■

The intuition behind this lemma is that: when $\{c_i\}_{i=1}^m$ are smooth and α_c-RLICQ holds for all $w \in \mathcal{W}$, then the constraint manifold has bounded curvature every where. Then, if we only care about first order behavior, it's well-approximated by the local dynamic in tangent plane, up to some second-order correction.

Therefore, by Eq.(112), we see locally it's not much different from the unconstrainted case Eq.(15) up to some negeligable correction. In the following analysis, we will always use formula Eq.(112) as the update equation for PSGD.

Since most of following proof bears a lot similarity as in unconstrained case, we only pointed out the essential steps in our following proof.

Theorem 36 (Main Theorem for Equality-Constrained Case) *Suppose a function $f(w) : \mathbb{R}^d \rightarrow \mathbb{R}$ with constraints $c_i(w) : \mathbb{R}^d \rightarrow \mathbb{R}$ is $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, and has a stochastic gradient oracle with radius at most Q, also satisfying $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi\xi^T = \sigma^2 I$. Further, suppose the function function f is B-bounded, L-Lipschitz, β-smooth, and has ρ-Lipschitz Hessian, and the constraints $\{c_i\}_{i=1}^m$ is L_i-Lipschitz, β_i-smooth, and has ρ_i-Lipschitz Hessian. Then there exists a threshold $\eta_{\max} = \tilde{\Theta}(1)$, so that for any $\zeta > 0$, and for any $\eta \leq \eta_{\max} / \max\{1, \log(1/\zeta)\}$, with probability at least $1 - \zeta$ in $t = \tilde{O}(\eta^{-2}\log(1/\zeta))$ iterations, PSGD outputs a point w_t that is $\tilde{O}(\sqrt{\eta \log(1/\eta\zeta)})$-close to some local minimum $w^\star$.*

First, we proof the assumptions in main theorem implies the smoothness conditions for $\mathfrak{M}(w)$, $\mathfrak{N}(w)$ and $\nabla^2_{ww}\mathcal{L}(w, \lambda^*(w'))$.

Lemma 37 *Under the assumptions of Theorem 36, there exists $\beta_M, \beta_N, \rho_M, \rho_N, \rho_L$ polynomial related to $B, L, \beta, \rho, \frac{1}{\alpha_c}$ and $\{L_i, \beta_i, \rho_i\}_{i=1}^m$ so that:*

1. $\|\mathfrak{M}(w)\| \leq \beta_M$ *and* $\|\mathfrak{N}(w)\| \leq \beta_N$ *for all* $w \in \mathcal{W}$.

2. $\mathfrak{M}(w)$ *is* ρ_M*-Lipschitz, and* $\mathfrak{N}(w)$ *is* ρ_N*-Lipschitz, and* $\nabla^2_{ww}\mathcal{L}(w, \lambda^*(w'))$ *is* ρ_L*-Lipschitz for all* $w' \in \mathcal{W}$.

Proof By definition of $\mathfrak{M}(w)$, $\mathfrak{N}(w)$ and $\nabla^2_{ww}\mathcal{L}(w, \lambda^*(w'))$, the above conditions will holds if there exists $B_\lambda, L_\lambda, \beta_\lambda$ bounded by $\tilde{O}(1)$, so that $\lambda^*(w)$ is B_λ-bounded, L_λ-Lipschitz, and β_λ-smooth.

By definition Eq.(82), we have:

$$\lambda^*(w) = C(w)^\dagger \nabla f(w) = (C(w)^T C(w))^{-1} C(w)^T \nabla f(w) \tag{113}$$

Because f is B-bounded, L-Lipschitz, β-smooth, and its Hessian is ρ-Lipschitz, thus, eventually, we only need to prove that there exists B_c, L_c, β_c bounded by $\tilde{O}(1)$, so that the pseudo-inverse $C(w)^\dagger$ is B_c-bounded, L_c-Lipschitz, and β_c-smooth.

Since α_c-RLICQ holds for all feasible points, we immediately have: $\|C(w)^\dagger\| \leq \frac{1}{\alpha_c}$, thus bounded. For simplicity, in the following context we use $C^\dagger$ to represent $C^\dagger(w)$ without ambiguity. By some calculation of linear algebra, we have the derivative of pseudo-inverse:

$$\frac{\partial C(w)^\dagger}{\partial w_i} = -C^\dagger \frac{\partial C(w)}{\partial w_i} C^\dagger + C^\dagger [C^\dagger]^T \frac{\partial C(w)^T}{\partial w_i}(I - CC^\dagger) \tag{114}$$

Again, α_c-RLICQ holds implies that derivative of pseudo-inverse is well-defined for every feasible point. Let tensor $E(w), \tilde{E}(w)$ to be the derivative of $C(w), C^\dagger(w)$, which is defined as:

$$[E(w)]_{ijk} = \frac{\partial [C(w)]_{ik}}{\partial w_j} \qquad [\tilde{E}(w)]_{ijk} = \frac{\partial [C(w)^\dagger]_{ik}}{\partial w_j} \tag{115}$$

Define the transpose of a 3rd order tensor $E^T_{i,j,k} = E_{k,j,i}$, then we have

$$\tilde{E}(w) = -[E(w)](C^\dagger, I, C^\dagger) + [E(w)^T](C^\dagger [C^\dagger]^T, I, (I - CC^\dagger)) \tag{116}$$

where by calculation $[E(w)](I, I, e_i) = \nabla^2 c_i(w)$.

Finally, since $C(w)^\dagger$ and $\nabla^2 c_i(w)$ are bounded by $\tilde{O}(1)$, by Eq.(116), we know $\tilde{E}(w)$ is bounded, that is $C(w)^\dagger$ is Lipschitz. Again, since both $C(w)^\dagger$ and $\nabla^2 c_i(w)$ are bounded, Lipschitz, by Eq.(116), we know $\tilde{E}(w)$ is also $\tilde{O}(1)$-Lipschitz. This finishes the proof. ∎

From now on, we can use the same proof strategy as unconstraint case. Below we list the corresponding lemmas and the essential steps that require modifications.

Lemma 38 *Under the assumptions of Theorem 36, with notations in Lemma 37, for any point with* $\|\chi(w_0)\| \geq \sqrt{2\eta\sigma^2\beta_M(d-m)}$ *where* $\sqrt{2\eta\sigma^2\beta_M(d-m)} < \epsilon$*, after one iteration we have:*

$$\mathbb{E}f(w_1) - f(w_0) \leq -\tilde{\Omega}(\eta^2) \tag{117}$$

Proof Choose $\eta_{\max} < \frac{1}{\beta_M}$, and also small enough, then by update equation Eq.(112), we have:

$$\begin{aligned}\mathbb{E}f(w_1) - f(w_0) &\leq \chi(w_0)^T\mathbb{E}(w_1 - w_0) + \frac{\beta_M}{2}\mathbb{E}\|w_1 - w_0\|^2 \\ &\leq -(\eta - \frac{\beta_M\eta^2}{2})\|\chi(w_0)\|^2 + \frac{\eta^2\sigma^2\beta_M(d-m)}{2} + \tilde{O}(\eta^2)\|\chi(w_0)\| + \tilde{O}(\eta^3) \\ &\leq -(\eta - \tilde{O}(\eta^{1.5}) - \frac{\beta_M\eta^2}{2})\|\chi(w_0)\|^2 + \frac{\eta^2\sigma^2\beta_M(d-m)}{2} + \tilde{O}(\eta^3) \\ &\leq -\frac{\eta^2\sigma^2\beta_M d}{4}\end{aligned} \tag{118}$$

Which finishes the proof. ∎

Theorem 39 *Under the assumptions of Theorem 36, with notations in Lemma 37, for any initial point w_0 that is $\tilde{O}(\sqrt{\eta}) < \delta$ close to a local minimum $w^\star$, with probability at least $1 - \zeta/2$, we have following holds simultaneously:*

$$\forall t \leq \tilde{O}(\frac{1}{\eta^2}\log\frac{1}{\zeta}), \quad \|w_t - w^\star\| \leq \tilde{O}(\sqrt{\eta\log\frac{1}{\eta\zeta}}) < \delta \tag{119}$$

where $w^\star$ is the locally optimal point.

Proof By calculus, we know

$$\chi(w_t) = \chi(w^\star) + \int_0^1 (\mathfrak{M} + \mathfrak{N})(w^\star + t(w_t - w^\star))\mathrm{d}t \cdot (w_t - w^\star) \tag{120}$$

Let filtration $\mathfrak{F}_t = \sigma\{\xi_0, \cdots \xi_{t-1}\}$, and note $\sigma\{\Delta_0, \cdots, \Delta_t\} \subset \mathfrak{F}_t$, where $\sigma\{\cdot\}$ denotes the sigma field. Let event $\mathfrak{E}_t = \{\forall\tau \leq t, \|w_\tau - w^\star\| \leq \mu\sqrt{\eta\log\frac{1}{\eta\zeta}} < \delta\}$, where μ is independent of (η, ζ), and will be specified later.

By Definition 34 of $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, we know $\mathfrak{M}(w)$ is locally α-strongly convex restricted to its tangent space $\mathcal{T}(w)$. in the 2δ-neighborhood of $w^\star$. If $\eta_{\max}$ is chosen small enough, by Remark 25 and Lemma 30, we have in addition:

$$\begin{aligned}\chi(w_t)^T(w_t - w^\star)1_{\mathfrak{E}_t} &= (w_t - w^\star)^T\int_0^1 (\mathfrak{M} + \mathfrak{N})(w^\star + t(w_t - w^\star))\mathrm{d}t \cdot (w_t - w^\star)1_{\mathfrak{E}_t} \\ &\geq [\alpha\|w_t - w^\star\|^2 - \tilde{O}(\|w_t - w^\star\|^3)]1_{\mathfrak{E}_t} \geq 0.5\alpha\|w_t - w^\star\|^2 1_{\mathfrak{E}_t}\end{aligned} \tag{121}$$

Then, everything else follows almost the same as the proof of Lemma 16. ∎

Lemma 40 *Under the assumptions of Theorem 36, with notations in Lemma 37, for any initial point w_0 where $\|\chi(w_0)\| \leq \tilde{O}(\eta) < \epsilon$, and $\hat{v}^T\mathfrak{M}(w_0)\hat{v} \leq -\gamma$ for some $\hat{v} \in \mathcal{T}(w)$, $\|\hat{v}\| = 1$, then there is a number of steps T that depends on w_0 such that:*

$$\mathbb{E}f(w_T) - f(w_0) \leq -\tilde{\Omega}(\eta) \tag{122}$$

The number of steps T has a fixed upper bound T_{max} that is independent of w_0 where $T \leq T_{max} = O((\log(d-m))/\gamma\eta)$.

Similar to the unconstrained case, we show this by a coupling sequence. Here the sequence we construct will only walk on the tangent space, by Lemmas in previous subsection, we know this is not very far from the actual sequence. We first define and characterize the coupled sequence in the following lemma:

Lemma 41 *Under the assumptions of Theorem 36, with notations in Lemma 37. Let $\tilde{f}$ defined as local second-order approximation of $f(x)$ around w_0 in tangent space $\mathcal{T}_0 = \mathcal{T}(w_0)$:*

$$\tilde{f}(w) \doteq f(w_0) + \chi(w_0)^T(w - w_0) + \frac{1}{2}(w - w_0)^T[P_{\mathcal{T}_0}^T \mathfrak{M}(w_0) P_{\mathcal{T}_0}](w - w_0) \tag{123}$$

$\{\tilde{w}_t\}$ be the corresponding sequence generated by running SGD on function $\tilde{f}$, with $\tilde{w}_0 = w_0$, and noise projected to $\mathcal{T}_0$, (i.e. $\tilde{w}_t = \tilde{w}_{t-1} - \eta(\tilde{\chi}(\tilde{w}_{t-1}) + P_{\mathcal{T}_0}\xi_{t-1})$. For simplicity, denote $\tilde{\chi}(w) = \nabla \tilde{f}(w)$, and $\widetilde{\mathfrak{M}} = P_{\mathcal{T}_0}^T \mathfrak{M}(w_0) P_{\mathcal{T}_0}$, then we have analytically:

$$\tilde{\chi}(\tilde{w}_t) = (1 - \eta\widetilde{\mathfrak{M}})^t \tilde{\chi}(\tilde{w}_0) - \eta\widetilde{\mathfrak{M}} \sum_{\tau=0}^{t-1} (1 - \eta\widetilde{\mathfrak{M}})^{t-\tau-1} P_{\mathcal{T}_0}\xi_\tau \tag{124}$$

$$\tilde{w}_t - w_0 = -\eta \sum_{\tau=0}^{t-1} (1 - \eta\widetilde{\mathfrak{M}})^\tau \tilde{\chi}(\tilde{w}_0) - \eta \sum_{\tau=0}^{t-1} (1 - \eta\widetilde{\mathfrak{M}})^{t-\tau-1} P_{\mathcal{T}_0}\xi_\tau \tag{125}$$

Furthermore, for any initial point w_0 where $\|\chi(w_0)\| \leq \tilde{O}(\eta) < \epsilon$, and $\min_{\hat{v}\in\mathcal{T}(w),\|\hat{v}\|=1} \hat{v}^T\mathfrak{M}(w_0)\hat{v} = -\gamma_0$. There exist a $T \in \mathbb{N}$ satisfying:

$$\frac{d - m}{\eta\gamma_0} \leq \sum_{\tau=0}^{T-1} (1 + \eta\gamma_0)^{2\tau} < \frac{3(d - m)}{\eta\gamma_0} \tag{126}$$

with probability at least $1 - \tilde{O}(\eta^3)$, we have following holds simultaneously for all $t \leq T$:

$$\|\tilde{w}_t - w_0\| \leq \tilde{O}(\eta^{\frac{1}{2}} \log \frac{1}{\eta}); \qquad \|\tilde{\chi}(\tilde{w}_t)\| \leq \tilde{O}(\eta^{\frac{1}{2}} \log \frac{1}{\eta}) \tag{127}$$

Proof Clearly we have:

$$\tilde{\chi}(\tilde{w}_t) = \tilde{\chi}(\tilde{w}_{t-1}) + \widetilde{\mathfrak{M}}(\tilde{w}_t - \tilde{w}_{t-1}) \tag{128}$$

and

$$\tilde{w}_t = \tilde{w}_{t-1} - \eta(\tilde{\chi}(\tilde{w}_{t-1}) + P_{\mathcal{T}_0}\xi_{t-1}) \tag{129}$$

This lemma is then proved by a direct application of Lemma 19. ■

Then we show the sequence constructed is very close to the actual sequence.

Lemma 42 *Under the assumptions of Theorem 36, with notations in Lemma 37. Let $\{w_t\}$ be the corresponding sequence generated by running PSGD on function f. Also let $\tilde{f}$ and $\{\tilde{w}_t\}$ be defined as in Lemma 41. Then, for any initial point w_0 where $\|\chi(w_0)\|^2 \leq \tilde{O}(\eta) < \epsilon$, and $\min_{\hat{v}\in\mathcal{T}(w),\|\hat{v}\|=1} \hat{v}^T\mathfrak{M}(w_0)\hat{v} = -\gamma_0$. Given the choice of T as in Eq.(126), with probability at least $1 - \tilde{O}(\eta^2)$, we have following holds simultaneously for all $t \leq T$:*

$$\|w_t - \tilde{w}_t\| \leq \tilde{O}(\eta \log^2 \frac{1}{\eta}); \tag{130}$$

Proof First, we have update function of tangent gradient by:

$$
\begin{aligned}
\chi(w_t) =& \chi(w_{t-1}) + \int_0^1 \nabla\chi(w_{t-1} + t(w_t - w_{t-1}))\mathrm{d}t \cdot (w_t - w_{t-1}) \\
=& \chi(w_{t-1}) + \mathfrak{M}(w_{t-1})(w_t - w_{t-1}) + \mathfrak{N}(w_{t-1})(w_t - w_{t-1}) + \theta_{t-1}
\end{aligned} \tag{131}
$$

where the remainder:

$$
\theta_{t-1} \equiv \int_0^1 \left[\nabla\chi(w_{t-1} + t(w_t - w_{t-1})) - \nabla\chi(w_{t-1})\right] \mathrm{d}t \cdot (w_t - w_{t-1}) \tag{132}
$$

Project it to tangent space $\mathcal{T}_0 = \mathcal{T}(w_0)$. Denote $\widetilde{\mathfrak{M}} = P_{\mathcal{T}_0}^T \mathfrak{M}(w_0) P_{\mathcal{T}_0}$, and $\widetilde{\mathfrak{M}}'_{t-1} = P_{\mathcal{T}_0}^T[\ \mathfrak{M}(w_{t_1}) - \mathfrak{M}(w_0)\ ]P_{\mathcal{T}_0}$. Then, we have:

$$
\begin{aligned}
P_{\mathcal{T}_0} \cdot \chi(w_t) =& P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) + P_{\mathcal{T}_0}(\mathfrak{M}(w_{t-1}) + \mathfrak{N}(w_{t-1}))(w_t - w_{t-1}) + P_{\mathcal{T}_0}\theta_{t-1} \\
=& P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0}(w_t - w_{t-1}) \\
& + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0^c}(w_t - w_{t-1}) + P_{\mathcal{T}_0}\mathfrak{N}(w_{t-1})(w_t - w_{t-1}) + P_{\mathcal{T}_0}\theta_{t-1} \\
=& P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) + \widetilde{\mathfrak{M}}(w_t - w_{t-1}) + \phi_{t-1}
\end{aligned} \tag{133}
$$

Where

$$
\phi_{t-1} = [\ \widetilde{\mathfrak{M}}'_{t-1} + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0^c} + P_{\mathcal{T}_0}\mathfrak{N}(w_{t-1})\ ](w_t - w_{t-1}) + P_{\mathcal{T}_0}\theta_{t-1} \tag{134}
$$

By Hessian smoothness, we immediately have:

$$
\|\widetilde{\mathfrak{M}}'_{t-1}\| = \|\mathfrak{M}(w_{t_1}) - \mathfrak{M}(w_0)\| \leq \rho_M\|w_{t-1} - w_0\| \leq \rho_M(\|w_t - \tilde{w}_t\| + \|\tilde{w}_t - w_0\|) \tag{135}
$$

$$
\|\theta_{t-1}\| \leq \frac{\rho_M + \rho_N}{2}\|w_t - w_{t-1}\|^2 \tag{136}
$$

Substitute the update equation of PSGD (Eq.(112)) into Eq.(133), we have:

$$
\begin{aligned}
& P_{\mathcal{T}_0} \cdot \chi(w_t) = P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) - \eta\widetilde{\mathfrak{M}}(P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) + P_{\mathcal{T}_0} \cdot P_{\mathcal{T}(w_{t-1})}\xi_{t-1}) + \widetilde{\mathfrak{M}} \cdot \iota_{t-1} + \phi_{t-1} \\
& = (1 - \eta\widetilde{\mathfrak{M}})P_{\mathcal{T}_0} \cdot \chi(w_{t-1}) - \eta\widetilde{\mathfrak{M}}P_{\mathcal{T}_0}\xi_{t-1} + \eta\widetilde{\mathfrak{M}}P_{\mathcal{T}_0} \cdot P_{\mathcal{T}^c(w_{t-1})}\xi_{t-1} + \widetilde{\mathfrak{M}} \cdot \iota_{t-1} + \phi_{t-1}
\end{aligned} \tag{137}
$$

Let $\Delta_t = P_{\mathcal{T}_0} \cdot \chi(w_t) - \tilde{\chi}(\tilde{w}_t)$ denote the difference of tangent gradient in $\mathcal{T}(w_0)$, then from Eq.(128), Eq.(129), and Eq.(137) we have:

$$
\Delta_t = (1 - \eta H)\Delta_{t-1} + \eta\widetilde{\mathfrak{M}}P_{\mathcal{T}_0} \cdot P_{\mathcal{T}^c(w_{t-1})}\xi_{t-1} + \widetilde{\mathfrak{M}} \cdot \iota_{t-1} + \phi_{t-1} \tag{138}
$$

$$
P_{\mathcal{T}_0} \cdot (w_t - w_0) - (\tilde{w}_t - w_0) = -\eta\sum_{\tau=0}^{t-1}\Delta_\tau + \eta\sum_{\tau=0}^{t-1}P_{\mathcal{T}_0} \cdot P_{\mathcal{T}^c(w_\tau)}\xi_\tau + \sum_{\tau=0}^{t-1}\iota_\tau \tag{139}
$$

By Lemma 30, we know if $\sum_{i=1}^m \frac{\beta_i^2}{\alpha_c^2} = \frac{1}{R^2}$, then we have:

$$
\|P_{\mathcal{T}_0^c}(w_t - w_0)\| \leq \frac{\|w_t - w_0\|^2}{2R} \tag{140}
$$

Let filtration $\mathfrak{F}_t = \sigma\{\xi_0, \cdots \xi_{t-1}\}$, and note $\sigma\{\Delta_0, \cdots, \Delta_t\} \subset \mathfrak{F}_t$, where $\sigma\{\cdot\}$ denotes the sigma field. Also, let event $\mathfrak{K}_t = \{\forall\tau \leq t,\ \|\tilde{\chi}(\tilde{w}_\tau)\| \leq \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta}),\ \|\tilde{w}_\tau - w_0\| \leq \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta})\}$, and denote

$\Gamma_t = \eta\sum_{\tau=0}^{t-1} P_{\mathcal{T}_0}\cdot P_{\mathcal{T}^c(w_\tau)}\xi_\tau$, let $\mathfrak{E}_t = \{\forall\tau \le t,\ \|\Delta_\tau\| \le \mu_1\eta\log^2\frac{1}{\eta}, \|\Gamma_\tau\| \le \mu_2\eta\log^2\frac{1}{\eta}, \|w_\tau - \tilde{w}_\tau\| \le \mu_3\eta\log^2\frac{1}{\eta}\}$ where (μ_1,μ_2,μ_3) are is independent of (η,ζ), and will be determined later. To prevent ambiguity in the proof, $\tilde{O}$ notation will not hide any dependence on μ. Clearly event $\mathfrak{K}_{t-1}\subset\mathfrak{F}_{t-1}, \mathfrak{E}_{t-1}\subset\mathfrak{F}_{t-1}$ thus independent of ξ_{t-1}.

Then, conditioned on event $\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}$, by triangle inequality, we have $\|w_\tau - w_0\| \le \tilde{O}(\eta^{\frac{1}{2}}\log\frac{1}{\eta})$, for all $\tau \le t-1 \le T-1$. We then need to carefully bound the following bound each term in Eq.(138). We know $w_t - w_{t-1} = -\eta\cdot(\chi(w_{t-1}) + P_{\mathcal{T}(w_{t-1})}\xi_{t-1}) + \iota_{t-1}$, and then by Lemma 32 and Lemma 31, we have:

$$\begin{aligned}
\|\eta\widetilde{\mathfrak{M}}P_{\mathcal{T}_0}\cdot P_{\mathcal{T}^c(w_{t-1})}\xi_{t-1}\| &\le \tilde{O}(\eta^{1.5}\log\frac{1}{\eta})\\
\|\widetilde{\mathfrak{M}}\cdot\iota_{t-1}\| &\le \tilde{O}(\eta^2)\\
\|[\,\widetilde{\mathfrak{M}}'_{t-1} + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0^c} + P_{\mathcal{T}_0}\mathfrak{N}(w_{t-1})\,](-\eta\cdot\chi(w_{t-1}))\| &\le \tilde{O}(\eta^2\log^2\frac{1}{\eta})\\
\|[\,\widetilde{\mathfrak{M}}'_{t-1} + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0^c} + P_{\mathcal{T}_0}\mathfrak{N}(w_{t-1})\,](-\eta P_{\mathcal{T}(w_{t-1})}\xi_{t-1})\| &\le \tilde{O}(\eta^{1.5}\log\frac{1}{\eta})\\
\|[\,\widetilde{\mathfrak{M}}'_{t-1} + P_{\mathcal{T}_0}\mathfrak{M}(w_{t-1})P_{\mathcal{T}_0^c} + P_{\mathcal{T}_0}\mathfrak{N}(w_{t-1})\,]\iota_{t-1}\| &\le \tilde{O}(\eta^2)\\
\|P_{\mathcal{T}_0}\theta_{t-1}\| &\le \tilde{O}(\eta^2)
\end{aligned}\tag{141}$$

Therefore, abstractly, conditioned on event $\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}$, we could write down the recursive equation as:

$$\Delta_t = (1-\eta H)\Delta_{t-1} + A + B \tag{142}$$

where $\|A\| \le \tilde{O}(\eta^{1.5}\log\frac{1}{\eta})$ and $\|B\| \le \tilde{O}(\eta^2\log^2\frac{1}{\eta})$, and in addition, by independence, easy to check we also have $\mathbb{E}[(1-\eta H)\Delta_{t-1}A|\mathfrak{F}_{t-1}] = 0$. This is exactly the same case as in the proof of Lemma 20. By the same argument of martingale and Azuma-Hoeffding, and by choosing μ_1 large enough, we can prove

$$P\left(\mathfrak{E}_{t-1}\cap\left\{\|\Delta_t\| \ge \mu_1\eta\log^2\frac{1}{\eta}\right\}\right) \le \tilde{O}(\eta^3) \tag{143}$$

On the other hand, for $\Gamma_t = \eta\sum_{\tau=0}^{t-1} P_{\mathcal{T}_0}\cdot P_{\mathcal{T}^c(w_\tau)}\xi_\tau$, we have:

$$\begin{aligned}
\mathbb{E}[\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}|\mathfrak{F}_{t-1}] &= \left[\Gamma_{t-1} + \eta\mathbb{E}[P_{\mathcal{T}_0}\cdot P_{\mathcal{T}^c(w_{t-1})}\xi_{t-1}|\mathfrak{F}_{t-1}]\right]1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}\\
&= \Gamma_{t-1}1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \le \Gamma_{t-1}1_{\mathfrak{K}_{t-2}\cap\mathfrak{E}_{t-2}}
\end{aligned}\tag{144}$$

Therefore, we have $\mathbb{E}[\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}] \le \Gamma_{t-1}1_{\mathfrak{K}_{t-2}\cap\mathfrak{E}_{t-2}}$ which means $\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}$ is a supermartingale.

We also know by Lemma 32, with probability 1:

$$\begin{aligned}
&|\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} - \mathbb{E}[\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \mid \mathfrak{F}_{t-1}]| = |\eta P_{\mathcal{T}_0}\cdot P_{\mathcal{T}^c(w_{t-1})}\xi_{t-1}|\cdot 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}}\\
\le&\tilde{O}(\eta)\|w_{t-1} - w_0\|1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} \le \tilde{O}(\eta^{1.5}\log\frac{1}{\eta}) = c_{t-1}
\end{aligned}\tag{145}$$

By Azuma-Hoeffding inequality, with probability less than $\tilde{O}(\eta^3)$, for $t \le T \le O(\log(d-m)/\gamma_0\eta)$:

$$\Gamma_t 1_{\mathfrak{K}_{t-1}\cap\mathfrak{E}_{t-1}} - \Gamma_0\cdot 1 > \tilde{O}(1)\sqrt{\sum_{\tau=0}^{t-1} c_\tau^2\log(\frac{1}{\eta})} = \tilde{O}(\eta\log^2\frac{1}{\eta}) \tag{146}$$

This means there exists some $\tilde{C}_2 = \tilde{O}(1)$ so that:

$$P\left(\mathfrak{K}_{t-1} \cap \mathfrak{E}_{t-1} \cap \left\{\|\Gamma_t\| \geq \tilde{C}_2 \eta \log^2 \frac{1}{\eta}\right\}\right) \leq \tilde{O}(\eta^3) \tag{147}$$

by choosing $\mu_2 > \tilde{C}_2$, we have:

$$P\left(\mathfrak{K}_{t-1} \cap \mathfrak{E}_{t-1} \cap \left\{\|\Gamma_t\| \geq \mu_2 \eta \log^2 \frac{1}{\eta}\right\}\right) \leq \tilde{O}(\eta^3) \tag{148}$$

Therefore, combined with Lemma 41, we have:

$$P\left(\mathfrak{E}_{t-1} \cap \left\{\|\Gamma_t\| \geq \mu_2 \eta \log^2 \frac{1}{\eta}\right\}\right) \leq \tilde{O}(\eta^3) + P(\overline{\mathfrak{K}}_{t-1}) \leq \tilde{O}(\eta^3) \tag{149}$$

Finally, conditioned on event $\mathfrak{K}_{t-1} \cap \mathfrak{E}_{t-1}$, if we have $\|\Gamma_t\| \leq \mu_2 \eta \log^2 \frac{1}{\eta}$, then by Eq.(139):

$$\|P_{\mathcal{T}_0} \cdot (w_t - w_0) - (\tilde{w}_t - w_0)\| \leq \tilde{O}\left((\mu_1 + \mu_2)\eta \log^2 \frac{1}{\eta}\right) \tag{150}$$

Since $\|w_{t-1} - w_0\| \leq \tilde{O}(\eta^{\frac{1}{2}} \log \frac{1}{\eta})$, and $\|w_t - w_{t-1}\| \leq \tilde{O}(\eta)$, by Eq.(140):

$$\|P_{\mathcal{T}_0^c}(w_t - w_0)\| \leq \frac{\|w_t - w_0\|^2}{2R} \leq \tilde{O}(\eta \log^2 \frac{1}{\eta}) \tag{151}$$

Thus:

$$\begin{aligned}\|w_t - \tilde{w}_t\|^2 =& \|P_{\mathcal{T}_0} \cdot (w_t - \tilde{w}_t)\|^2 + \|P_{\mathcal{T}_0^c} \cdot (w_t - \tilde{w}_t)\|^2 \\ =& \|P_{\mathcal{T}_0} \cdot (w_t - w_0) - (\tilde{w}_t - w_0)\|^2 + \|P_{\mathcal{T}_0^c}(w_t - w_0)\|^2 \leq \tilde{O}((\mu_1 + \mu_2)^2 \eta^2 \log^4 \frac{1}{\eta})\end{aligned} \tag{152}$$

That is there exist some $\tilde{C}_3 = \tilde{O}(1)$ so that $\|w_t - \tilde{w}_t\| \leq \tilde{C}_3(\mu_1 + \mu_2)\eta \log^2 \frac{1}{\eta}$ Therefore, conditioned on event $\mathfrak{K}_{t-1} \cap \mathfrak{E}_{t-1}$, we have proved that if choose $\mu_3 > \tilde{C}_3(\mu_1 + \mu_2)$, then event $\{\|w_t - \tilde{w}_t\| \geq \mu_3 \eta \log^2 \frac{1}{\eta}\} \subset \{\|\Gamma_t\| \geq \mu_2 \eta \log^2 \frac{1}{\eta}\}$. Then, combined this fact with Eq.(143), Eq.(149), we have proved:

$$P\left(\mathfrak{E}_{t-1} \cap \overline{\mathfrak{E}}_t\right) \leq \tilde{O}(\eta^3) \tag{153}$$

Because $P(\overline{\mathfrak{E}}_0) = 0$, and $T \leq \tilde{O}(\frac{1}{\eta})$, we have $P(\overline{\mathfrak{E}}_T) \leq \tilde{O}(\eta^2)$, which concludes the proof.

■

These two lemmas allow us to prove the result when the initial point is very close to a saddle point.

Proof [Proof of Lemma 40] Combine Talyor expansion Eq.87 with Lemma 41, Lemma 42, we prove this Lemma by the same argument as in the proof of Lemma 17. ■

Finally the main theorem follows.

Proof [Proof of Theorem 36] By Lemma 38, Lemma 40, and Lemma 39, with the same argument as in the proof Theorem 14, we easily concludes this proof. ■

Appendix C. Detailed Proofs for Section 4

In this section we show two optimization problems (11) and (13) satisfy the $(\alpha, \gamma, \epsilon, \delta)$-strict saddle propery.

C.1. Warm up: maximum eigenvalue formulation

Recall that we are trying to solve the optimization (11), which we restate here.

$$\begin{aligned} &\max \quad T(u,u,u,u), \\ &\|u\|^2 = 1. \end{aligned} \tag{154}$$

Here the tensor T has orthogonal decomposition $T = \sum_{i=1}^{d} a_i^{\otimes 4}$. We first do a change of coordinates to work in the coordinate system specified by (a_i)'s (this does not change the dynamics of the algorithm). In particular, let $u = \sum_{i=1}^{d} x_i a_i$ (where $x \in \mathbb{R}^d$), then we can see $T(u,u,u,u) = \sum_{i=1}^{d} x_i^4$. Therefore let $f(x) = -\|x\|_4^4$, the optimization problem is equivalent to

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & \|x\|_2^2 = 1 \end{aligned} \tag{155}$$

This is a constrained optimization, so we apply the framework developed in Section 3.3.
Let $c(x) = \|x\|_2^2 - 1$. We first compute the Lagrangian

$$\mathcal{L}(x,\lambda) = f(x) - \lambda c(x) = -\|x\|_4^4 - \lambda(\|x\|_2^2 - 1). \tag{156}$$

Since there is only one constraint, and the gradient when $\|x\| = 1$ always have norm 2, we know the set of constraints satisfy 2-RLICQ. In particular, we can compute the correct value of Lagrangian multiplier λ,

$$\lambda^*(x) = \arg\min_{\lambda} \|\nabla_x \mathcal{L}(x,\lambda)\| = \arg\min_{\lambda} \sum_{i=1}^{d} (2x_i^3 + \lambda x_i)^2 = -2\|x\|_4^4 \tag{157}$$

Therefore, the gradient in the tangent space is equal to

$$\begin{aligned} \chi(x) &= \nabla_x \mathcal{L}(x,\lambda)|_{(x,\lambda^*(x))} = \nabla f(x) - \lambda^*(x)\nabla c(x) \\ &= -4(x_1^3, \cdots, x_d^3)^T - 2\lambda^*(x)(x_1, \cdots, x_d)^T \\ &= 4\left((x_1^2 - \|x\|_4^4)x_1, \cdots, (x_d^2 - \|x\|_4^4)x_d\right) \end{aligned} \tag{158}$$

The second-order partial derivative of Lagrangian is equal to

$$\begin{aligned} \mathfrak{M}(x) &= \nabla_{xx}^2 \mathcal{L}(x,\lambda)|_{(x,\lambda^*(x))} = \nabla^2 f(x) - \lambda^*(x)\nabla^2 c(x) \\ &= -12\text{diag}(x_1^2, \cdots, x_d^2) - 2\lambda^*(x) I_d \\ &= -12\text{diag}(x_1^2, \cdots, x_d^2) + 4\|x\|_4^4 I_d \end{aligned} \tag{159}$$

Since the variable x has bounded norm, and the function is a polynomial, it's clear that the function itself is bounded and all its derivatives are bounded. Moreover, all the derivatives of the constraint are bounded. We summarize this in the following lemma.

Lemma 43 *The objective function (11) is bounded by* 1*, its* p*-th order derivative is bounded by* $O(\sqrt{d})$ *for* $p = 1, 2, 3$*. The constraint's* p*-th order derivative is bounded by* 2*, for* $p = 1, 2, 3$*.*

Therefore the function satisfy all the smoothness condition we need. Finally we show the gradient and Hessian of Lagrangian satisfy the $(\alpha, \gamma, \epsilon, \delta)$-strict saddle property. Note that we did not try to optimize the dependency with respect to d.

Theorem 44 *The only local minima of optimization problem (11) are $\pm a_i$ ($i \in [d]$). Further it satisfy $(\alpha, \gamma, \epsilon, \delta)$-strict saddle for $\gamma = 7/d$, $\alpha = 3$ and $\epsilon, \delta = 1/poly(d)$.*

In order to prove this theorem, we consider the transformed version Eq.155. We first need following two lemma for points around saddle point and local minimum respectively. We choose

$$\epsilon_0 = (10d)^{-4}, \;\; \epsilon = 4\epsilon_0^2, \;\; \delta = 2d\epsilon_0, \;\; \mathfrak{S}(x) = \{i \mid |x_i| > \epsilon_0\} \tag{160}$$

Where by intuition, $\mathfrak{S}(x)$ is the set of coordinates whose value is relative large.

Lemma 45 *Under the choice of parameters in Eq.(160), suppose $\|\chi(x)\| \le \epsilon$, and $|\mathfrak{S}(x)| \ge 2$. Then, there exists $\hat{v} \in \mathcal{T}(x)$ and $\|\hat{v}\| = 1$, so that $\hat{v}^T \mathfrak{M}(x) \hat{v} \le -7/d$.*

Proof Suppose $|\mathfrak{S}(x)| = p$, and $2 \le p \le d$. Since $\|\chi(x)\| \le \epsilon = 4\epsilon_0^2$, by Eq.(158), we have for each $i \in [d]$, $|[\chi(x)]_i| = 4|(x_i^2 - \|x\|_4^4)x_i| \le 4\epsilon_0^2$. Therefore, we have:

$$\forall i \in \mathfrak{S}(x), \qquad |x_i^2 - \|x\|_4^4| \le \epsilon_0 \tag{161}$$

and thus:

$$\begin{aligned} &||\|x\|_4^4 - \frac{1}{p}| = ||\|x\|_4^4 - \frac{1}{p}\sum_i x_i^2| \\ \le &||\|x\|_4^4 - \frac{1}{p}\sum_{i\in\mathfrak{S}(x)} x_i^2| + |\frac{1}{p}\sum_{i\in[d]-\mathfrak{S}(x)} x_i^2| \le \epsilon_0 + \frac{d-p}{p}\epsilon_0^2 \le 2\epsilon_0 \end{aligned} \tag{162}$$

Combined with Eq.161, this means:

$$\forall i \in \mathfrak{S}(x), \qquad |x_i^2 - \frac{1}{p}| \le 3\epsilon_0 \tag{163}$$

Because of symmetry, WLOG we assume $\mathfrak{S}(x) = \{1, \cdots, p\}$. Since $|\mathfrak{S}(x)| \ge 2$, we can pick $\hat{v} = (a, b, 0, \cdots, 0)$. Here $a > 0, b < 0$, and $a^2 + b^2 = 1$. We pick a such that $ax_1 + bx_2 = 0$. The solution is the intersection of a radius 1 circle and a line which passes $(0, 0)$, which always exists. For this $\hat{v}$, we know $\|\hat{v}\| = 1$, and $\hat{v}^T x = 0$ thus $\hat{v} \in \mathcal{T}(x)$. We have:

$$\begin{aligned} &\hat{v}^T \mathfrak{M}(x) \hat{v} = -(12x_1^2 + 4\|x\|_4^4)a^2 - (12x_2^2 + 4\|x\|_4^4)b^2 \\ = &-8x_1^2a^2 - 8x_2^2b^2 - 4(x_1^2 - \|x\|_4^4))a^2 - 4(x_2^2 - \|x\|_4^4))b^2 \\ \le &-\frac{8}{p} + 24\epsilon_0 + 4\epsilon_0 \le -7/d \end{aligned} \tag{164}$$

Which finishes the proof. ■

Lemma 46 *Under the choice of parameters in Eq.(160), suppose $\|\chi(x)\| \le \epsilon$, and $|\mathfrak{S}(x)| = 1$. Then, there is a local minimum $x^\star$ such that $\|x - x^\star\| \le \delta$, and for all x' in the 2δ neighborhood of $x^\star$, we have $\hat{v}^T \mathfrak{M}(x')\hat{v} \ge 3$ for all $\hat{v} \in \mathcal{T}(x')$, $\|\hat{v}\| = 1$*

Proof WLOG, we assume $\mathfrak{S}(x) = \{1\}$. Then, we immediately have for all $i > 1$, $|x_i| \le \epsilon_0$, and thus:

$$1 \ge x_1^2 = 1 - \sum_{i>1} x_i^2 \ge 1 - d\epsilon_0^2 \tag{165}$$

Therefore $x_1 \ge \sqrt{1 - d\epsilon_0^2}$ or $x_1 \le -\sqrt{1 - d\epsilon_0^2}$. Which means x_1 is either close to 1 or close to -1. By symmetry, we know WLOG, we can assume the case $x_1 \ge \sqrt{1 - d\epsilon_0^2}$. Let $e_1 = (1, 0, \cdots, 0)$, then we know:

$$\|x - e_1\|^2 \le (x_1 - 1)^2 + \sum_{i>1} x_i^2 \le 2d\epsilon_0^2 \le \delta^2 \tag{166}$$

Next, we show e_1 is a local minimum. According to Eq.159, we know $\mathfrak{M}(e_1)$ is a diagonal matrix with 4 on the diagonals except for the first diagonal entry (which is equal to -8), since $\mathcal{T}(e_1) = \text{span}\{e_2, \cdots, e_d\}$, we have:

$$v^T \mathfrak{M}(e_1) v \ge 4\|v\|^2 > 0 \qquad \text{for all } v \in \mathcal{T}(e_1), v \ne 0 \tag{167}$$

Which by Theorem 28 means e_1 is a local minimum.

Finally, denote $\mathcal{T}_1 = \mathcal{T}(e_1)$ be the tangent space of constraint manifold at e_1. We know for all x' in the 2δ neighborhood of e_1, and for all $\hat{v} \in \mathcal{T}(x')$, $\|\hat{v}\| = 1$:

$$\begin{aligned}
\hat{v}^T \mathfrak{M}(x')\hat{v} \ge& \hat{v}^T \mathfrak{M}(e_1)\hat{v} - |\hat{v}^T \mathfrak{M}(e_1)\hat{v} - \hat{v}^T \mathfrak{M}(x')\hat{v}| \\
=& 4\|P_{\mathcal{T}_1}\hat{v}\|^2 - 8\|P_{\mathcal{T}_1^c}\hat{v}\|^2 - \|\mathfrak{M}(e_1) - \mathfrak{M}(x')\|\|\hat{v}\|^2 \\
=& 4 - 12\|P_{\mathcal{T}_1^c}\hat{v}\|^2 - \|\mathfrak{M}(e_1) - \mathfrak{M}(x')\|
\end{aligned} \tag{168}$$

By lemma 31, we know $\|P_{\mathcal{T}_1^c}\hat{v}\|^2 \le \|x' - e_1\|^2 \le 4\delta^2$. By Eq.(159), we have:

$$\begin{aligned}
&\|\mathfrak{M}(e_1) - \mathfrak{M}(x')\| \le \|\mathfrak{M}(e_1) - \mathfrak{M}(x')\| \le \sum_{(i,j)} |[\mathfrak{M}(e_1)]_{ij} - [\mathfrak{M}(x')]_{ij}| \\
\le& \sum_i \left| -12[e_1]_i^2 + 4\|e_1\|_4^4 - 12x_i^2 + 4\|x\|_4^4 \right| \le 64d\delta
\end{aligned} \tag{169}$$

In conclusion, we have $\hat{v}^T \mathfrak{M}(x')\hat{v} \ge 4 - 48\delta^2 - 64d\delta \ge 3$ which finishes the proof. ∎

Finally, we are ready to prove Theorem 44.

Proof [Proof of Theorem 44]

According to Lemma 45 and Lemma 46, we immediately know the optimization problem satisfies $(\alpha, \gamma, \epsilon, \delta)$-strict saddle.

The only thing remains to show is that the only local minima of optimization problem (11) are $\pm a_i$ ($i \in [d]$). Which is equivalent to show that the only local minima of the transformed problem is $\pm e_i$ ($i \in [d]$), where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$, where 1 is on i-th coordinate.

By investigating the proof of Lemma 45 and Lemma 46, we know these two lemmas actually hold for any small enough choice of ϵ_0 satisfying $\epsilon_0 \le (10d)^{-4}$, by pushing $\epsilon_0 \to 0$, we know for any point satisfying $|\chi(x)| \le \epsilon \to 0$, if it is close to some local minimum, it must satisfy $1 = |\mathfrak{S}(x)| \to \text{supp}(x)$. Therefore, we know the only possible local minima are $\pm e_i$ ($i \in [d]$). In Lemma 46, we proved e_1 is local minimum, by symmetry, we finishes the proof. ■

C.2. New formulation

In this section we consider our new formulation (13). We first restate the optimization problem here:

$$\begin{aligned} \min \quad & \sum_{i \neq j} T(u^{(i)}, u^{(i)}, u^{(j)}, u^{(j)}), \\ \forall i \quad & \|u^{(i)}\|^2 = 1. \end{aligned} \tag{170}$$

Note that we changed the notation for the variables from u_i to $u^{(i)}$, because in later proofs we will often refer to the particular coordinates of these vectors.

Similar to the previous section, we perform a change of basis. The effect is equivalent to making a_i's equal to basis vectors e_i (and hence the tensor is equal to $T = \sum_{i=1}^{d} e_i^{\otimes 4}$. After the transformation the equations become

$$\begin{aligned} \min \quad & \sum_{(i,j):i \neq j} h(u^{(i)}, u^{(j)}) \\ \text{s.t.} \quad & \|u^{(i)}\|^2 = 1 \qquad \forall i \in [d] \end{aligned} \tag{171}$$

Here $h(u^{(i)}, u^{(j)}) = \sum_{k=1}^{d} (u_k^{(i)} u_k^{(j)})^2$, $(i, j) \in [d]^2$. We divided the objective function by 2 to simplify the calculation.

Let $U \in \mathbb{R}^{d^2}$ be the concatenation of $\{u^{(i)}\}$ such that $U_{ij} = u_j^{(i)}$. Let $c_i(U) = \|u^{(i)}\|^2 - 1$ and $f(U) = \frac{1}{2} \sum_{(i,j):i \neq j} h(u^{(i)}, u^{(j)})$. We can then compute the Lagrangian

$$\mathcal{L}(U, \lambda) = f(U) - \sum_{i=1}^{d} \lambda_i c_i(U) = \frac{1}{2} \sum_{(i,j):i \neq j} h(u^{(i)}, u^{(j)}) - \sum_{i=1}^{d} \lambda_i (\|u^{(i)}\|^2 - 1) \tag{172}$$

The gradients of $c_i(U)$'s are equal to $(0, \cdots, 0, 2u^{(i)}, 0, \cdots, 0)^T$, all of these vectors are orthogonal to each other (because they have disjoint supports) and have norm 2. Therefore the set of constraints satisfy 2-RLICQ. We can then compute the Lagrangian multipiers λ^* as follows

$$\lambda^*(U) = \arg\min_{\lambda} \|\nabla_U \mathcal{L}(U, \lambda)\| = \arg\min_{\lambda} 4 \sum_i \sum_k (\sum_{j:j \neq i} U_{jk}^2 U_{ik} - \lambda_i U_{ik})^2 \tag{173}$$

which gives:

$$\lambda_i^*(U) = \arg\min_{\lambda} \sum_k (\sum_{j:j \neq i} U_{jk}^2 U_{ik} - \lambda_i U_{ik})^2 = \sum_{j:j \neq i} h(u^{(j)}, u^{(i)}) \tag{174}$$

Therefore, gradient in the tangent space is equal to

$$\chi(U) = \nabla_U \mathcal{L}(U, \lambda)|_{(U, \lambda^*(U))} = \nabla f(U) - \sum_{i=1}^{n} \lambda_i^*(U) \nabla c_i(U). \tag{175}$$

The gradient is a d^2 dimensional vector (which can be viewed as a $d \times d$ matrix corresponding to entries of U), and we express this in a coordinate-by-coordinate way. For simplicity of later proof, denote:

$$\psi_{ik}(U) = \sum_{j:j\neq i}[U_{jk}^2 - h(u^{(j)}, u^{(i)})] = \sum_{j:j\neq i}[U_{jk}^2 - \sum_{l=1}^{d} U_{il}^2 U_{jl}^2] \tag{176}$$

Then we have:

$$\begin{aligned}
[\chi(U)]_{ik} &= 2(\sum_{j:j\neq i} U_{jk}^2 - \lambda_i^*(U))U_{ik} \\
&= 2U_{ik}\sum_{j:j\neq i}(U_{jk}^2 - h(u^{(j)}, u^{(i)})) \\
&= 2U_{ik}\psi_{ik}(U)
\end{aligned} \tag{177}$$

Similarly we can compute the second-order partial derivative of Lagrangian as

$$\mathfrak{M}(U) = \nabla^2 f(U) - \sum_{i=1}^{d} \lambda_i^* \nabla^2 c_i(U). \tag{178}$$

The Hessian is a $d^2 \times d^2$ matrix, we index it by 4 indices in $[d]$. The entries are summarized below:

$$\begin{aligned}
[\mathfrak{M}(U)]_{ik,i'k'} &= \frac{\partial}{\partial U_{i'k'}}[\nabla_U \mathcal{L}(U,\lambda)]_{ik}\Big|_{(U,\lambda^*(U))} = \frac{\partial}{\partial U_{i'k'}}[2(\sum_{j:j\neq i} U_{jk}^2 - \lambda)U_{ik}]\Big|_{(U,\lambda^*(U))} \\
&= \begin{cases} 2(\sum_{j:j\neq i} U_{jk}^2 - \lambda_i^*(U)) & \text{if } k = k', i = i' \\ 4U_{i'k}U_{ik} & \text{if } k = k', i \neq i' \\ 0 & \text{if } k \neq k' \end{cases} \\
&= \begin{cases} 2\psi_{ik}(U) & \text{if } k = k', i = i' \\ 4U_{i'k}U_{ik} & \text{if } k = k', i \neq i' \\ 0 & \text{if } k \neq k' \end{cases}
\end{aligned} \tag{179}$$

Similar to the previous case, it is easy to bound the function value and derivatives of the function and the constraints.

Lemma 47 *The objective function (13) and p-th order derivative are all bounded by poly(d) for $p = 1, 2, 3$. Each constraint's p-th order derivative is bounded by 2, for $p = 1, 2, 3$.*

Therefore the function satisfy all the smoothness condition we need. Finally we show the gradient and Hessian of Lagrangian satisfy the $(\alpha, \gamma, \epsilon, \delta)$-strict saddle property. Again we did not try to optimize the dependency with respect to d.

Theorem 48 *Optimization problem (13) has exactly $2^d \cdot d!$ local minimum that corresponds to permutation and sign flips of a_i's. Further, it satisfy $(\alpha, \gamma, \epsilon, \delta)$-strict saddle for $\alpha = 1$ and $\gamma, \epsilon, \delta = 1/poly(d)$.*

Again, in order to prove this theorem, we follow the same strategy: we consider the transformed version Eq.171. and first prove the following lemmas for points around saddle point and local minimum respectively. We choose

$$\epsilon_0 = (10d)^{-6}, \;\; \epsilon = 2\epsilon_0^6, \;\; \delta = 2d\epsilon_0, \;\; \gamma = \epsilon_0^4/4, \;\; \mathfrak{S}(u) = \{k \mid |u_k| > \epsilon_0\} \tag{180}$$

Where by intuition, $\mathfrak{S}(u)$ is the set of coordinates whose value is relative large.

Lemma 49 *Under the choice of parameters in Eq.(180), suppose $\|\chi(U)\| \le \epsilon$, and there exists $(i,j) \in [d]^2$ so that $\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)}) \ne \emptyset$. Then, there exists $\hat{v} \in \mathcal{T}(U)$ and $\|\hat{v}\| = 1$, so that $\hat{v}^T \mathfrak{M}(U) \hat{v} \le -\gamma$.*

Proof Again, since $\|\chi(x)\| \le \epsilon = 2\epsilon_0^6$, by Eq.(177), we have for each $i \in [d]$, $|[\chi(x)]_{ik}| = 2|U_{ik}\psi_{ik}(U)| \le 2\epsilon_0^6$. Therefore, have:

$$\forall k \in \mathfrak{S}(u^{(i)}), \qquad |\psi_{ik}(U)| \le \epsilon_0^5 \tag{181}$$

Then, we prove this lemma by dividing it into three cases. Note in order to prove that there exists $\hat{v} \in \mathcal{T}(U)$ and $\|\hat{v}\| = 1$, so that $\hat{v}^T \mathfrak{M}(U) \hat{v} \le -\gamma$; it suffices to find a vector $v \in \mathcal{T}(U)$ and $\|v\| \le 1$, so that $v^T \mathfrak{M}(U) v \le -\gamma$.

Case 1 : $|\mathfrak{S}(u^{(i)})| \ge 2$, $|\mathfrak{S}(u^{(j)})| \ge 2$, and $|\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)})| \ge 2$.

WLOG, assume $\{1,2\} \in \mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)})$, choose v to be $v_{i1} = \frac{U_{i2}}{4}$, $v_{i2} = -\frac{U_{i1}}{4}$, $v_{j1} = \frac{U_{j2}}{4}$ and $v_{j2} = -\frac{U_{j1}}{4}$. All other entries of v are zero. Clearly $v \in \mathcal{T}(U)$, and $\|v\| \le 1$. On the other hand, we know $\mathfrak{M}(U)$ restricted to these 4 coordinates $(i1, i2, j1, j2)$ is

$$\begin{pmatrix} 2\psi_{i1}(U) & 0 & 4U_{i1}U_{j1} & 0 \\ 0 & 2\psi_{i2}(U) & 0 & 4U_{i2}U_{j2} \\ 4U_{i1}U_{j1} & 0 & 2\psi_{j1}(U) & 0 \\ 0 & 4U_{i2}U_{j2} & 0 & 2\psi_{j2}(U) \end{pmatrix} \tag{182}$$

By Eq.(181), we know all diagonal entries are $\le 2\epsilon_0^5$.

If $U_{i1}U_{j1}U_{i2}U_{j2}$ is negative, we have the quadratic form:

$$\begin{aligned} v^T \mathfrak{M}(U) v =& U_{i1}U_{j1}U_{i2}U_{j2} + \frac{1}{8}[U_{i2}^2\psi_{i1}(U) + U_{i1}^2\psi_{i2}(U) + U_{j2}^2\psi_{j1}(U) + U_{j1}^2\psi_{j2}(U)] \\ \le & -\epsilon_0^4 + \epsilon_0^5 \le -\frac{1}{4}\epsilon_0^4 = -\gamma \end{aligned} \tag{183}$$

If $U_{i1}U_{j1}U_{i2}U_{j2}$ is positive we just swap the sign of the first two coordinates $v_{i1} = -\frac{U_{i2}}{2}$, $v_{i2} = \frac{U_{i1}}{2}$ and the above argument would still holds.

Case 2 : $|\mathfrak{S}(u^{(i)})| \ge 2$, $|\mathfrak{S}(u^{(j)})| \ge 2$, and $|\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)})| = 1$.

WLOG, assume $\{1,2\} \in \mathfrak{S}(u^{(i)})$ and $\{1,3\} \in \mathfrak{S}(u^{(j)})$, choose v to be $v_{i1} = \frac{U_{i2}}{4}$, $v_{i2} = -\frac{U_{i1}}{4}$, $v_{j1} = \frac{U_{j3}}{4}$ and $v_{j3} = -\frac{U_{j1}}{4}$. All other entries of v are zero. Clearly $v \in \mathcal{T}(U)$ and $\|v\| \le 1$. On the other hand, we know $\mathfrak{M}(U)$ restricted to these 4 coordinates $(i1, i2, j1, j3)$ is

$$\begin{pmatrix} 2\psi_{i1}(U) & 0 & 4U_{i1}U_{j1} & 0 \\ 0 & 2\psi_{i2}(U) & 0 & 0 \\ 4U_{i1}U_{j1} & 0 & 2\psi_{j1}(U) & 0 \\ 0 & 0 & 0 & 2\psi_{j3}(U) \end{pmatrix} \tag{184}$$

By Eq.(181), we know all diagonal entries are $\leq 2\epsilon_0^5$. If $U_{i1}U_{j1}U_{i2}U_{j3}$ is negative, we have the quadratic form:

$$\begin{aligned} v^T\mathfrak{M}(U)v =& \frac{1}{2}U_{i1}U_{j1}U_{i2}U_{j3} + \frac{1}{8}[U_{i2}^2\psi_{i1}(U) + U_{i1}^2\psi_{i2}(U) + U_{j3}^2\psi_{j1}(U) + U_{j1}^2\psi_{j3}(U)] \\ \leq& -\frac{1}{2}\epsilon_0^4 + \epsilon_0^5 \leq -\frac{1}{4}\epsilon_0^4 = -\gamma \end{aligned} \tag{185}$$

If $U_{i1}U_{j1}U_{i2}U_{j3}$ is positive we just swap the sign of the first two coordinates $v_{i1} = -\frac{U_{i2}}{2}$, $v_{i2} = \frac{U_{i1}}{2}$ and the above argument would still holds.

Case 3 : Either $|\mathfrak{S}(u^{(i)})| = 1$ or $|\mathfrak{S}(u^{(j)})| = 1$.

WLOG, suppose $|\mathfrak{S}(u^{(i)})| = 1$, and $\{1\} = \mathfrak{S}(u^{(i)})$, we know:

$$|(u_1^{(i)})^2 - 1| \leq (d-1)\epsilon_0^2 \tag{186}$$

On the other hand, since $\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)}) \neq \emptyset$, we have $\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)}) = \{1\}$, and thus:

$$|\psi_{j1}(U)| = |\sum_{i':i'\neq j} U_{i'1}^2 - \sum_{i':i'\neq j} h(u^{(i')}, u^{(j)})| \leq \epsilon_0^5 \tag{187}$$

Therefore, we have:

$$\sum_{i':i'\neq j} h(u^{(i')}, u^{(j)}) \geq \sum_{i':i'\neq j} U_{i'1}^2 - \epsilon_0^5 \geq U_{i1}^2 - \epsilon_0^5 \geq 1 - d\epsilon_0^2 \tag{188}$$

and

$$\begin{aligned} \sum_{k=1}^d \psi_{jk}(U) =& \sum_{i':i'\neq j}\sum_{k=1}^d U_{i'k}^2 - d\sum_{i':i'\neq j} h(u^{(i')}, u^{(j)}) \\ \leq& d - 1 - d(1 - d\epsilon_0^2) = -1 + d^2\epsilon_0^2 \end{aligned} \tag{189}$$

Thus, we know, there must exist some $k' \in [d]$, so that $\psi_{jk'}(U) \leq -\frac{1}{d} + d\epsilon_0^2$. This means we have "large" negative entry on the diagonal of $\mathfrak{M}$. Since $|\psi_{j1}(U)| \leq \epsilon_0^5$, we know $k' \neq 1$. WLOG, suppose $k' = 2$, we have $|\psi_{j2}(U)| > \epsilon_0^5$, thus $|U_{j2}| \leq \epsilon_0$.

Choose v to be $v_{j1} = \frac{U_{j2}}{2}$, $v_{j2} = -\frac{U_{j1}}{2}$. All other entries of v are zero. Clearly $v \in \mathcal{T}(U)$ and $\|v\| \leq 1$. On the other hand, we know $\mathfrak{M}(U)$ restricted to these 2 coordinates $(j1, j2)$ is

$$\begin{pmatrix} 2\psi_{j1}(U) & 0 \\ 0 & 2\psi_{j2}(U) \end{pmatrix} \tag{190}$$

We know $|U_{j1}| > \epsilon_0$, $|U_{j2}| \leq \epsilon_0$, $|\psi_{j1}(U)| \leq \epsilon_0^5$, and $\psi_{j2}(U) \leq -\frac{1}{d} + d\epsilon_0^2$. Thus:

$$\begin{aligned} v^T\mathfrak{M}(U)v =& \frac{1}{2}\psi_{j1}(U)U_{j2}^2 + \frac{1}{2}\psi_{j2}(U)U_{j1}^2 \\ \leq& \epsilon_0^7 - (\frac{1}{d} - d\epsilon_0^2)\epsilon_0^2 \leq -\frac{1}{2d}\epsilon_0^2 \leq -\gamma \end{aligned} \tag{191}$$

Since by our choice of v, we have $\|v\| \leq 1$, we can choose $\hat{v} = v/\|v\|$, and immediately have $\hat{v} \in \mathcal{T}(U)$ and $\|\hat{v}\| = 1$, and $\hat{v}^T\mathfrak{M}(U)\hat{v} \leq -\gamma$. ∎

Lemma 50 *Under the choice of parameters in Eq.(180), suppose $\|\chi(U)\| \le \epsilon$, and for any $(i,j) \in [d]^2$ we have $\mathfrak{S}(u^{(i)}) \cap \mathfrak{S}(u^{(j)}) = \emptyset$. Then, there is a local minimum $U^\star$ such that $\|U - U^\star\| \le \delta$, and for all U' in the 2δ neighborhood of $U^\star$, we have $\hat{v}^T \mathfrak{M}(U')\hat{v} \ge 1$ for all $\hat{v} \in \mathcal{T}(U')$, $\|\hat{v}\| = 1$*

Proof WLOG, we assume $\mathfrak{S}(u^{(i)}) = \{i\}$ for $i = 1, \cdots, d$. Then, we immediately have:

$$|u_j^{(i)}| \le \epsilon_0, \qquad |(u_i^{(i)})^2 - 1| \le (d-1)\epsilon_0^2, \qquad \forall (i,j) \in [d]^2, j \ne i \tag{192}$$

Then $u_i^{(i)} \ge \sqrt{1 - d\epsilon_0^2}$ or $u_i^{(i)} \le -\sqrt{1 - d\epsilon_0^2}$. Which means $u_i^{(i)}$ is either close to 1 or close to -1. By symmetry, we know WLOG, we can assume the case $u_i^{(i)} \ge \sqrt{1 - d\epsilon_0^2}$ for all $i \in [d]$.

Let $V \in \mathbb{R}^{d^2}$ be the concatenation of $\{e_1, e_2, \cdots, e_d\}$, then we have:

$$\|U - V\|^2 = \sum_{i=1}^{d} \|u^{(i)} - e_i\|^2 \le 2d^2\epsilon_0^2 \le \delta^2 \tag{193}$$

Next, we show V is a local minimum. According to Eq.179, we know $\mathfrak{M}(V)$ is a diagonal matrix with d^2 entries:

$$[\mathfrak{M}(V)]_{ik,ik} = 2\psi_{ik}(V) = 2\sum_{j:j\ne i}[V_{jk}^2 - \sum_{l=1}^{d} V_{il}^2 V_{jl}^2] = \begin{cases} 2 & \text{if } i \ne k \\ 0 & \text{if } i = k \end{cases} \tag{194}$$

We know the unit vector in the direction that corresponds to $[\mathfrak{M}(V)]_{ii,ii}$ is not in the tangent space $\mathcal{T}(V)$ for all $i \in [d]$. Therefore, for any $v \in \mathcal{T}(V)$, we have

$$v^T \mathfrak{M}(e_1) v \ge 2\|v\|^2 > 0 \qquad \text{for all } v \in \mathcal{T}(V), v \ne 0 \tag{195}$$

Which by Theorem 28 means V is a local minimum.

Finally, denote $\mathcal{T}_V = \mathcal{T}(V)$ be the tangent space of constraint manifold at V. We know for all U' in the 2δ neighborhood of V, and for all $\hat{v} \in \mathcal{T}(x')$, $\|\hat{v}\| = 1$:

$$\begin{aligned}
\hat{v}^T \mathfrak{M}(U')\hat{v} \ge & \hat{v}^T \mathfrak{M}(V)\hat{v} - |\hat{v}^T \mathfrak{M}(V)\hat{v} - \hat{v}^T \mathfrak{M}(U')\hat{v}| \\
= & 2\|P_{\mathcal{T}_V}\hat{v}\|^2 - \|\mathfrak{M}(V) - \mathfrak{M}(U')\|\|\hat{v}\|^2 \\
= & 2 - 2\|P_{\mathcal{T}_V^c}\hat{v}\|^2 - \|\mathfrak{M}(V) - \mathfrak{M}(U')\|
\end{aligned} \tag{196}$$

By lemma 31, we know $\|P_{\mathcal{T}_V^c}\hat{v}\|^2 \le \|U' - V\|^2 \le 4\delta^2$. By Eq.(179), we have:

$$\|\mathfrak{M}(V) - \mathfrak{M}(U')\| \le \|\mathfrak{M}(V) - \mathfrak{M}(U')\| \le \sum_{(i,j,k)} |[\mathfrak{M}(V)]_{ik,jk} - [\mathfrak{M}(U')]_{ik,jk}| \le 100d^3\delta \tag{197}$$

In conclusion, we have $\hat{v}^T \mathfrak{M}(U')\hat{v} \ge 2 - 8\delta^2 - 100d^3\delta \ge 1$ which finishs the proof. ■

Finally, we are ready to prove Theorem 48.

Proof [Proof of Theorem 48]

Similarly, $(\alpha, \gamma, \epsilon, \delta)$-strict saddleimmediately follows from Lemma 49 and Lemma 50.

The only thing remains to show is that Optimization problem (13) has exactly $2^d \cdot d!$ local minimum that corresponds to permutation and sign flips of a_i's. This can be easily proved by the same argument as in the proof of Theorem 44. ■

C.3. Extending to tensors of different order

In this section we show how to generalize our algorithm to tensors of different orders. As a 8-th order tensor (and more generally, $4p$-th order tensor for $p \in \mathcal{N}^+$) can always be considered to be a 4-th order tensor with components $a_i^{\otimes} a_i$ ($a_i^{\otimes p}$ in general), so it is trivial to generalize our algorithm to 8-th order or any $4p$-th order.

For tensors of other orders, we need to apply some transformation. As a concrete example, we show how to transform an orthogonal 3rd order tensor into an orthogonal 4-th order tensor.

We first need to define a few notations. For third order tensors $A, B \in \mathbb{R}^{d^3}$, we define $(A \otimes B)_{i_1,i_2,\ldots,i_6} = A_{i_1,i_2,i_3} B_{i_4,i_5,i_6} (i_1, \ldots, i_6 \in [d])$. We also define the *partial trace* operation that maps a 6-th order tensor $T \in \mathbb{R}^{d^6}$ to a 4-th order tensor in $\mathbb{R}^{d^4}$:

$$ptrace(T)_{i_1,i_2,i_3,i_4} = \sum_{i=1}^{d} T(i, i_1, i_2, i, i_3, i_4).$$

Basically, the operation views the tensor as a $d^3 \times d^3$ matrix with $d^2 \times d^2$ $d \times d$ matrix blocks, then takes the trace of each matrix block. Now given a random variable $X \in \mathbb{R}^{d^3}$ whose expectation is an orthogonal third order tensor, we can use these operations to construct an orthogonal 4-th order tensor:

Lemma 51 *Suppose the expectation of random variable $X \in \mathbb{R}^{d^3}$ is an orthogonal 3rd order tensor:*

$$\mathbb{E}[X] = \sum_{i=1}^{d} a_i^{\otimes 3},$$

where a_i's are orthonormal vectors. Let X' be an independent sample of X, then we know

$$\mathbb{E}[ptrace(X \otimes X')] = \sum_{i=1}^{d} a_i^{\otimes 3}.$$

In other words, we can construct random samples whose expectation is equal to a 4-th order orthogonal tensor.

Proof Since $ptrace$ and $\otimes$ are all linear operations, by linearity of expectation we know

$$\mathbb{E}[ptrace(X \otimes X')] = ptrace(\mathbb{E}[X] \otimes \mathbb{E}[X']) = ptrace((\sum_{i=1}^{d} a_i^{\otimes 3}) \otimes (\sum_{i=1}^{d} a_i^{\otimes 3})).$$

We can then expand out the product:

$$(\sum_{i=1}^{d} a_i^{\otimes 3}) \otimes (\sum_{i=1}^{d} a_i^{\otimes 3}) = \sum_{i=1}^{d} a_i^{\otimes 6} + \sum_{i \neq j} a_i^{\otimes 3} \otimes a_j^{\otimes 3}.$$

For the diagonal terms, we know $ptrace(a_i^{\otimes} 6) = \|a_i\|^2 a_i^{\otimes} 4 = a_i^{\otimes} 4$. For the $i \neq j$ terms, we know $ptrace(a_i^{\otimes 3} \otimes a_j^{\otimes 3}) = \langle a_i, a_j \rangle a_i^{\otimes} 2 \otimes a_j^{\otimes} 2 = 0$ (since a_i, a_j are orthogonal). Therefore we must have

$$ptrace((\sum_{i=1}^{d} a_i^{\otimes 3}) \otimes (\sum_{i=1}^{d} a_i^{\otimes 3})) = \sum_{i=1}^{d} ptrace(a_i^{\otimes 6}) + \sum_{i \neq j} ptrace(a_i^{\otimes 3} \otimes a_j^{\otimes 3}) = \sum_{i=1}^{d} a_i^{\otimes 4}.$$

This gives the result. ■

Using similar operations we can easily convert all odd-order tensors into order $4p(p \in \mathbb{N}^+)$. For tensors of order $4p + 2(p \in \mathbb{N}^+)$, we can simply apply the partial trace and get a tensor of order $4p$ with desirable properties. Therefore our results applies for all orders of tensors.