A. Detailed Proof of Main Theorem

In this section, we give detailed proof for the main theorem. We will first state two key lemmas that show how the algorithm can make progress when the gradient is large or near a saddle point, and show how the main theorem follows from the two lemmas. Then we will focus on the novel technique in this paper: how to analyze gradient descent near saddle point.

A.1. General Framework

In order to prove the main theorem, we need to show that the algorithm will not be stuck at any point that either has a large gradient or is a saddle point. This idea is similar to previous works (e.g.(Ge et al., 2015)). We first state a standard Lemma that shows if the current gradient is large, then we make progress in function value.

Lemma 12. *Assume $f(\cdot)$ satisfies A1, then for gradient descent with stepsize $\eta < \frac{1}{\ell}$, we have:*

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2$$

Proof. By Assumption A1 and its property, we have:

$$\begin{aligned} f(\mathbf{x}_{t+1}) \le& f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{\ell}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ =& f(\mathbf{x}_t) - \eta\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta^2\ell}{2}\|\nabla f(\mathbf{x}_t)\|^2 \le f(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2 \end{aligned}$$

□

The next lemma says that if we are "close to a saddle points", i.e., we are at a point where the gradient is small, but the Hessian has a reasonably large negative eigenvalue. This is the main difficulty in the analysis. We show a perturbation followed by small (polylog) number of standard gradient descent steps can also make the function value decrease with high probability.

Lemma 13. *There exist absolute constant $c_{\max}$, for $f(\cdot)$ satisfies A1, and any $c \le c_{\max}$, and $\chi \ge 1$. Let $\eta, r, g_{thres}, f_{thres}, t_{thres}$ calculated same way as in Algorithm 2. Then, if $\tilde{\mathbf{x}}_t$ satisfies:*

$$\|\nabla f(\tilde{\mathbf{x}}_t)\| \le g_{thres} \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}}_t)) \le -\sqrt{\rho\epsilon}$$

Let $\mathbf{x}_t = \tilde{\mathbf{x}}_t + \xi_t$ where ξ_t comes from the uniform distribution over $\mathbb{B}_0(r)$, and let $\mathbf{x}_{t+i}$ be the iterates of gradient descent from $\mathbf{x}_t$ with stepsize η, then with at least probability $1 - \frac{d\ell}{\sqrt{\rho\epsilon}}e^{-\chi}$, we have:

$$f(\mathbf{x}_{t+t_{thres}}) - f(\tilde{\mathbf{x}}_t) \le -f_{thres}$$

The proof of this lemma is deferred to Section A.2. Using this Lemma, we can then prove the main Theorem.

Theorem 3. *There exist absolute constant $c_{\max}$ such that: if $f(\cdot)$ satisfies A1, then for any $\delta > 0, \epsilon \le \frac{\ell^2}{\rho}, \Delta_f \ge f(\mathbf{x}_0) - f^\star$, and constant $c \le c_{\max}$, with probability $1-\delta$, the output of PGD$(\mathbf{x}_0, \ell, \rho, \epsilon, c, \delta, \Delta_f)$ will be $\epsilon-$second order stationary point, and terminate in iterations:*

$$O\left(\frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}\log^4\left(\frac{d\ell\Delta_f}{\epsilon^2\delta}\right)\right)$$

Proof. Denote $\tilde{c}_{\max}$ to be the absolute constant allowed in Theorem 13. In this theorem, we let $c_{\max} = \min\{\tilde{c}_{\max}, 1/2\}$, and choose any constant $c \le c_{\max}$.

In this proof, we will actually achieve some point satisfying following condition:

$$\|\nabla f(\mathbf{x})\| \le g_{\text{thres}} = \frac{\sqrt{c}}{\chi^2}\cdot\epsilon, \qquad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\sqrt{\rho\epsilon} \tag{3}$$

Since $c \le 1$, $\chi \ge 1$, we have $\frac{\sqrt{c}}{\chi^2} \le 1$, which implies any $\mathbf{x}$ satisfy Eq.(3) is also a ϵ-second-order stationary point.

Starting from $\mathbf{x}_0$, we know if $\mathbf{x}_0$ does not satisfy Eq.(3), there are only two possibilities:

1. $\|\nabla f(\mathbf{x}_0)\| > g_{\text{thres}}$: In this case, Algorithm 2 will not add perturbation. By Lemma 12:

$$f(\mathbf{x}_1) - f(\mathbf{x}_0) \le -\frac{\eta}{2} \cdot g_{\text{thres}}^2 = -\frac{c^2}{2\chi^4} \cdot \frac{\epsilon^2}{\ell}$$

2. $\|\nabla f(\mathbf{x}_0)\| \le g_{\text{thres}}$: In this case, Algorithm 2 will add a perturbation of radius r, and will perform gradient descent (without perturbations) for the next t_{thres} steps. Algorithm 2 will then check termination condition. If the condition is not met, we must have:

$$f(\mathbf{x}_{t_{\text{thres}}}) - f(\mathbf{x}_0) \le -f_{\text{thres}} = -\frac{c}{\chi^3} \cdot \sqrt{\frac{\epsilon^3}{\rho}}$$

This means on average every step decreases the function value by

$$\frac{f(\mathbf{x}_{t_{\text{thres}}}) - f(\mathbf{x}_0)}{t_{\text{thres}}} \le -\frac{c^3}{\chi^4} \cdot \frac{\epsilon^2}{\ell}$$

In case 1, we can repeat this argument for $t = 1$ and in case 2, we can repeat this argument for $t = t_{\text{thres}}$. Hence, we can conclude as long as algorithm 2 has not terminated yet, on average, every step decrease function value by at least $\frac{c^3}{\chi^4} \cdot \frac{\epsilon^2}{\ell}$. However, we clearly can not decrease function value by more than $f(\mathbf{x}_0) - f^\star$, where $f^\star$ is the function value of global minima. This means algorithm 2 must terminate within the following number of iterations:

$$\frac{f(\mathbf{x}_0) - f^\star}{\frac{c^3}{\chi^4} \cdot \frac{\epsilon^2}{\ell}} = \frac{\chi^4}{c^3} \cdot \frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} = O\left(\frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} \log^4\left(\frac{d\ell\Delta_f}{\epsilon^2\delta}\right)\right)$$

Finally, we would like to ensure when Algorithm 2 terminates, the point it finds is actually an ϵ-second-order stationary point. The algorithm can only terminate when the gradient is small, and the function value does not decrease after a perturbation and t_{thres} iterations. We shall show every time when we add perturbation to iterate $\tilde{\mathbf{x}}_t$, if $\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}}_t)) < -\sqrt{\rho\epsilon}$, then we will have $f(\mathbf{x}_{t+t_{\text{thres}}}) - f(\tilde{\mathbf{x}}_t) \le -f_{\text{thres}}$. Thus, whenever the current point is not an ϵ-second-order stationary point, the algorithm cannot terminate.

According to Algorithm 2, we immediately know $\|\nabla f(\tilde{\mathbf{x}}_t)\| \le g_{\text{thres}}$ (otherwise we will not add perturbation at time t). By Lemma 13, we know this event happens with probability at least $1 - \frac{d\ell}{\sqrt{\rho\epsilon}} e^{-\chi}$ each time. On the other hand, during one entire run of Algorithm 2, the number of times we add perturbations is at most:

$$\frac{1}{t_{\text{thres}}} \cdot \frac{\chi^4}{c^3} \cdot \frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} = \frac{\chi^3}{c} \frac{\sqrt{\rho\epsilon}(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}$$

By union bound, for all these perturbations, with high probability Lemma 13 is satisfied. As a result Algorithm 2 works correctly. The probability of that is at least

$$1 - \frac{d\ell}{\sqrt{\rho\epsilon}} e^{-\chi} \cdot \frac{\chi^3}{c} \frac{\sqrt{\rho\epsilon}(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} = 1 - \frac{\chi^3 e^{-\chi}}{c} \cdot \frac{d\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}$$

Recall our choice of $\chi = 3\max\{\log(\frac{d\ell\Delta_f}{c\epsilon^2\delta}), 4\}$. Since $\chi \ge 12$, we have $\chi^3 e^{-\chi} \le e^{-\chi/3}$, this gives:

$$\frac{\chi^3 e^{-\chi}}{c} \cdot \frac{d\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2} \le e^{-\chi/3} \frac{d\ell(f(\mathbf{x}_0) - f^\star)}{c\epsilon^2} \le \delta$$

which finishes the proof.

□

A.2. Main Lemma: Escaping from Saddle Points Quickly

Now we prove the main lemma, which shows near a saddle point, a small perturbation followed by a small number of gradient descent steps will decrease the function value with high probability. This is the main step where we need new analysis, as the analysis previous works (e.g.(Ge et al., 2015)) do not work when the step size and perturbation do not depend polynomially in dimension d.

Intuitively, after adding a perturbation, the current point of the algorithm is a uniform distribution over a d-dimensional ball centered at $\tilde{\mathbf{x}}$, which we call **perturbation ball**. After a small number of gradient steps, some points in this ball (which we call the **escaping region**) will significantly decrease the function; other points (which we call the **stuck region**) does not see a significant decrease in function value. We hope to show that the escaping region constitutes at least $1-\delta$ fraction of the volume of the perturbation ball.

However, we do not know the exact form of the function near the saddle point, so the escaping region does not have a clean analytic description. Explicitly computing its volume can be very difficult. Our proof rely on a crucial observation: although we do not know the shape of the stuck region, we know the "width" of it must be small, therefore it cannot have a large volume. We will formalize this intuition later in Lemma 15.

The proof of the main lemma requires carefully balancing between different quantities including function value, gradient, parameter space and number of iterations. For clarify, we define following scalar quantities, which serve as the "units" for function value, gradient, parameter space, and time (iterations). We will use these notations throughout the proof.

Let the condition number be the ratio of the smoothness parameter (largest eigenvalue of Hessian) and the negative eigenvalue γ: $\kappa = \ell/\gamma \geq 1$, we define the following units:

$$\mathscr{F} := \eta\ell\frac{\gamma^3}{\rho^2}\cdot\log^{-3}(\frac{d\kappa}{\delta}), \qquad \mathscr{G} := \sqrt{\eta\ell}\frac{\gamma^2}{\rho}\cdot\log^{-2}(\frac{d\kappa}{\delta})$$

$$\mathscr{S} := \sqrt{\eta\ell}\frac{\gamma}{\rho}\cdot\log^{-1}(\frac{d\kappa}{\delta}), \qquad \mathscr{T} := \frac{\log(\frac{d\kappa}{\delta})}{\eta\gamma}$$

Intuitively, if we plug in our choice of step size $\eta\ell = O(1)$ (which we will prove later) and hide the logarithmic dependences, we have $\mathscr{F} = \tilde{O}(\frac{\gamma^3}{\rho^2}), \mathscr{G} = \tilde{O}(\frac{\gamma^2}{\rho}), \mathscr{S} = \tilde{O}(\frac{\gamma}{\rho})$, which is the only way to correctly discribe the units of function value, gradient, parameter space by just γ and ρ. Moreover, these units are closely related, in particular, we know $\sqrt{\frac{\mathscr{F}\cdot\log(\frac{d\kappa}{\delta})}{\gamma}} = \frac{\mathscr{G}\cdot\log(\frac{d\kappa}{\delta})}{\gamma} = \mathscr{S}$.

For simplicity of later proofs, we first restate Lemma 13 into a slightly more general form as follows. Lemma 13 is directly implied following lemma.

Lemma 14 (Lemma 13 restated)**.** *There exists universal constant $c_{\max}$, for $f(\cdot)$ satisfies A1, for any $\delta \in (0, \frac{d\kappa}{e}]$, suppose we start with point $\tilde{\mathbf{x}}$ satisfying following conditions:*

$$\|\nabla f(\tilde{\mathbf{x}})\| \leq \mathscr{G} \qquad \textit{and} \qquad \lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}})) \leq -\gamma$$

Let $\mathbf{x}_0 = \tilde{\mathbf{x}} + \xi$ where ξ come from the uniform distribution over ball with radius $\mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta}))$, and let $\mathbf{x}_t$ be the iterates of gradient descent from $\mathbf{x}_0$. Then, when stepsize $\eta \leq c_{\max}/\ell$, with at least probability $1-\delta$, we have following for any $T \geq \frac{1}{c_{\max}}\mathscr{T}$:

$$f(\mathbf{x}_T) - f(\tilde{\mathbf{x}}) \leq -\mathscr{F}$$

Lemma 14 is almost the same as Lemma 13. It is easy to verify that by substituting $\eta = \frac{c}{\ell}, \gamma = \sqrt{\rho\epsilon}$ and $\delta = \frac{d\ell}{\sqrt{\rho\epsilon}}e^{-\chi}$ into Lemma 14, we immediately obtain Lemma 13.

Now we will formalize the intuition that the "width" of stuck region is small.

Lemma 15. *There exists a universal constant $c_{\max}$, for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the conditions in Lemma 14, and without loss of generality let $\mathbf{e}_1$ be the minimum eigenvector of $\nabla^2 f(\tilde{\mathbf{x}})$. Consider two gradient descent sequences $\{\mathbf{u}_t\}, \{\mathbf{w}_t\}$ with initial points $\mathbf{u}_0, \mathbf{w}_0$ satisfying: (denote radius $r = \mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta}))$)*

$$\|\mathbf{u}_0 - \tilde{\mathbf{x}}\| \leq r, \quad \mathbf{w}_0 = \mathbf{u}_0 + \mu\cdot r\cdot\mathbf{e}_1, \quad \mu \in [\delta/(2\sqrt{d}), 1]$$

Then, for any stepsize $\eta \leq c_{\max}/\ell$, and any $T \geq \frac{1}{c_{\max}}\mathscr{T}$, we have:

$$\min\{f(\mathbf{u}_T) - f(\mathbf{u}_0), f(\mathbf{w}_T) - f(\mathbf{w}_0)\} \leq -2.5\mathscr{F}$$

Intuitively, lemma 15 claims for any two points $\mathbf{u}_0, \mathbf{w}_0$ inside the perturbation ball, if $\mathbf{u}_0 - \mathbf{w}_0$ lies in the direction of minimum eigenvector of $\nabla^2 f(\tilde{\mathbf{x}})$, and $\|\mathbf{u}_0 - \mathbf{w}_0\|$ is greater than threshold $\delta r/(2\sqrt{d})$, then at least one of two sequences $\{\mathbf{u}_t\}, \{\mathbf{w}_t\}$ will "efficiently escape saddle point". In other words, if $\mathbf{u}_0$ is a point in the stuck region, consider any point $\mathbf{w}_0$ that is on a straight line along direction of $\mathbf{e}_1$. As long as $\mathbf{w}_0$ is slightly far ($\delta r/\sqrt{d}$) from $\mathbf{u}_0$, it must be in the escaping region. This is what we mean by the "width" of the stuck region being small.Now we prove the main Lemma using this observation:

Proof of Lemma 14. By adding perturbation, in worst case we increase function value by:

$$f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}) \leq \nabla f(\tilde{\mathbf{x}})^\top \xi + \frac{\ell}{2}\|\xi\|^2 \leq \mathscr{G}(\frac{\mathscr{S}}{\kappa \cdot \log(\frac{d\kappa}{\delta})}) + \frac{1}{2}\ell(\frac{\mathscr{S}}{\kappa \cdot \log(\frac{d\kappa}{\delta})})^2 \leq \frac{3}{2}\mathscr{F}$$

On the other hand, let radius $r = \frac{\mathscr{S}}{\kappa \cdot \log(\frac{d\kappa}{\delta})}$. We know $\mathbf{x}_0$ come froms uniform distribution over $\mathbb{B}_{\tilde{\mathbf{x}}}(r)$. Let $\mathcal{X}_{\text{stuck}} \subset \mathbb{B}_{\tilde{\mathbf{x}}}(r)$ denote the set of bad starting points so that if $\mathbf{x}_0 \in \mathcal{X}_{\text{stuck}}$, then $f(\mathbf{x}_T) - f(\mathbf{x}_0) > -2.5\mathscr{F}$ (thus stuck at a saddle point); otherwise if $\mathbf{x}_0 \in B_{\tilde{\mathbf{x}}}(r) - \mathcal{X}_{\text{stuck}}$, we have $f(\mathbf{x}_T) - f(\mathbf{x}_0) \leq -2.5\mathscr{F}$.

By applying Lemma 15, we know for any $\mathbf{x}_0 \in \mathcal{X}_{\text{stuck}}$, it is guaranteed that $(\mathbf{x}_0 \pm \mu r \mathbf{e}_1) \not\in \mathcal{X}_{\text{stuck}}$ where $\mu \in [\frac{\delta}{2\sqrt{d}}, 1]$. Denote $I_{\mathcal{X}_{\text{stuck}}}(\cdot)$ be the indicator function of being inside set $\mathcal{X}_{\text{stuck}}$; and vector $\mathbf{x} = (x^{(1)}, \mathbf{x}^{(-1)})$, where $x^{(1)}$ is the component along $\mathbf{e}_1$ direction, and $\mathbf{x}^{(-1)}$ is the remaining $d-1$ dimensional vector. Recall $\mathbb{B}^{(d)}(r)$ be d-dimensional ball with radius r; By calculus, this gives an upper bound on the volumn of $\mathcal{X}_{\text{stuck}}$:

$$\begin{aligned}
\text{Vol}(\mathcal{X}_{\text{stuck}}) &= \int_{\mathbb{B}^{(d)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x} \cdot I_{\mathcal{X}_{\text{stuck}}}(\mathbf{x}) \\
&= \int_{\mathbb{B}^{(d-1)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x}^{(-1)} \int_{\tilde{x}^{(1)} - \sqrt{r^2 - \|\tilde{\mathbf{x}}^{(-1)} - \mathbf{x}^{(-1)}\|^2}}^{\tilde{x}^{(1)} + \sqrt{r^2 - \|\tilde{\mathbf{x}}^{(-1)} - \mathbf{x}^{(-1)}\|^2}} \mathrm{d}x^{(1)} \cdot I_{\mathcal{X}_{\text{stuck}}}(\mathbf{x}) \\
&\leq \int_{\mathbb{B}^{(d-1)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x}^{(-1)} \cdot \left(2 \cdot \frac{\delta}{2\sqrt{d}} r\right) = \text{Vol}(\mathbb{B}_0^{(d-1)}(r)) \times \frac{\delta r}{\sqrt{d}}
\end{aligned}$$

Then, we immediately have the ratio:

$$\frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(\mathbb{B}^{(d)}_{\tilde{\mathbf{x}}}(r))} \leq \frac{\frac{\delta r}{\sqrt{d}} \times \text{Vol}(\mathbb{B}_0^{(d-1)}(r))}{\text{Vol}(\mathbb{B}_0^{(d)}(r))} = \frac{\delta}{\sqrt{\pi d}} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{\delta}{\sqrt{\pi d}} \cdot \sqrt{\frac{d}{2} + \frac{1}{2}} \leq \delta$$

The second last inequality is by the property of Gamma function that $\frac{\Gamma(x+1)}{\Gamma(x+1/2)} < \sqrt{x + \frac{1}{2}}$ as long as $x \geq 0$. Therefore, with at least probability $1 - \delta$, $\mathbf{x}_0 \not\in \mathcal{X}_{\text{stuck}}$. In this case, we have:

$$\begin{aligned}
f(\mathbf{x}_T) - f(\tilde{\mathbf{x}}) =& f(\mathbf{x}_T) - f(\mathbf{x}_0) + f(\mathbf{x}_0) - f(0) \\
\leq& -2.5\mathscr{F} + 1.5\mathscr{F} \leq -\mathscr{F}
\end{aligned}$$

which finishes the proof. □

A.3. Bounding the Width of Stuck Region

In order to prove Lemma 15, we do it in two steps:

1. We first show if gradient descent from $\mathbf{u}_0$ does not decrease function value, then all the iterates must lie within a small ball around $\mathbf{u}_0$ (Lemma 16).
2. If gradient descent starting from a point $\mathbf{u}_0$ stuck in a small ball around a saddle point, then gradient descent from $\mathbf{w}_0$ (moving $\mathbf{u}_0$ along $\mathbf{e}_1$ direction for at least a certain distance), will decreases the function value (Lemma 17).

Recall we assumed without loss of generality $\mathbf{e}_1$ is the minimum eigenvector of $\nabla^2 f(\tilde{\mathbf{x}})$. In this context, we denote $\mathcal{H} := \nabla^2 f(\tilde{\mathbf{x}})$, and for simplicity of calculation, we consider following quadratic approximation:

$$\tilde{f}_{\mathbf{y}}(\mathbf{x}) := f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \mathcal{H} (\mathbf{x} - \mathbf{y}) \tag{4}$$

Now we are ready to state two lemmas formally:

Lemma 16. *For any constant $\hat{c} \geq 3$, there exists absolute constant $c_{\max}$: for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the condition in Lemma 14, for any initial point $\mathbf{u}_0$ with $\|\mathbf{u}_0 - \tilde{\mathbf{x}}\| \leq 2\mathscr{S}/(\kappa \cdot \log(\frac{d\kappa}{\delta}))$, define:*

$$T = \min\left\{ \inf_t \left\{t | \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_t) - f(\mathbf{u}_0) \leq -3\mathscr{F}\right\}, \hat{c}\mathscr{T} \right\}$$

then, for any $\eta \leq c_{\max}/\ell$, we have for all $t < T$ that $\|\mathbf{u}_t - \tilde{\mathbf{x}}\| \leq 100(\mathscr{S} \cdot \hat{c})$.

Lemma 17. *There exists absolute constant $c_{\max}, \hat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the condition in Lemma 14, and sequences $\{\mathbf{u}_t\}, \{\mathbf{w}_t\}$ satisfy the conditions in Lemma 15, define:*

$$T = \min\left\{ \inf_t \left\{t | \tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) - f(\mathbf{w}_0) \leq -3\mathscr{F}\right\}, \hat{c}\mathscr{T} \right\}$$

then, for any $\eta \leq c_{\max}/\ell$, if $\|\mathbf{u}_t - \tilde{\mathbf{x}}\| \leq 100(\mathscr{S} \cdot \hat{c})$ for all $t < T$, we will have $T < \hat{c}\mathscr{T}$.

Note the conclusion $T < \hat{c}\mathscr{T}$ in Lemma 17 equivalently means:

$$\inf_t \left\{t | \tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) - f(\mathbf{w}_0) \leq -3\mathscr{F}\right\} < \hat{c}\mathscr{T}$$

That is, for some $T < \hat{c}\mathscr{T}$, $\{\mathbf{w}_t\}$ sequence "escape the saddle point" in the sense of sufficient function value decrease $\tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) - f(\mathbf{w}_0) \leq -3\mathscr{F}$. Now, we are ready to prove Lemma 15.

Proof of Lemma 15. W.L.O.G, let $\tilde{\mathbf{x}} = 0$ be the origin. Let $(c^{(2)}_{\max}, \hat{c})$ be the absolute constant so that Lemma 17 holds, also let $c^{(1)}_{\max}$ be the absolute constant to make Lemma 16 holds based on our current choice of $\hat{c}$. We choose $c_{\max} \leq \min\{c^{(1)}_{\max}, c^{(2)}_{\max}\}$ so that our step size $\eta \leq c_{\max}/\ell$ is small enough which make both Lemma 16 and Lemma 17 hold. Let $T^\star := \hat{c}\mathscr{T}$ and define:

$$T' = \inf_t \left\{t | \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_t) - f(\mathbf{u}_0) \leq -3\mathscr{F}\right\}$$

Let's consider following two cases:

Case $T' \leq T^\star$: In this case, by Lemma 16, we know $\|\mathbf{u}_{T'-1}\| \leq O(\mathscr{S})$, and therefore

$$\|\mathbf{u}_{T'}\| \leq \|\mathbf{u}_{T'-1}\| + \eta\|\nabla f(\mathbf{u}_{T'-1})\| \leq \|\mathbf{u}_{T'-1}\| + \eta\|\nabla f(\tilde{\mathbf{x}})\| + \eta\ell\|\mathbf{u}_{T'-1}\| \leq O(\mathscr{S})$$

By choosing $c_{\max}$ small enough and $\eta \leq c_{\max}/\ell$, this gives:

$$\begin{aligned} f(\mathbf{u}_{T'}) - f(\mathbf{u}_0) \leq& \nabla f(\mathbf{u}_0)^\top(\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{1}{2}(\mathbf{u}_{T'} - \mathbf{u}_0)^\top \nabla^2 f(\mathbf{u}_0)(\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{\rho}{6}\|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 \\ \leq& \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_t) - f(\mathbf{u}_0) + \frac{\rho}{2}\|\mathbf{u}_0 - \tilde{\mathbf{x}}\|\|\mathbf{u}_{T'} - \mathbf{u}_0\|^2 + \frac{\rho}{6}\|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 \\ \leq& -3\mathscr{F} + O(\rho\mathscr{S}^3) = -3\mathscr{F} + O(\sqrt{\eta\ell} \cdot \mathscr{F}) \leq -2.5\mathscr{F} \end{aligned}$$

By choose $c_{\max} \leq \min\{1, \frac{1}{\hat{c}}\}$. We know $\eta < \frac{1}{\ell}$, by Lemma 12, we know gradient descent always decrease function value. Therefore, for any $T \geq \frac{1}{c_{\max}}\mathscr{T} \geq \hat{c}\mathscr{T} = T^\star \geq T'$, we have:

$$f(\mathbf{u}_T) - f(\mathbf{u}_0) \leq f(\mathbf{u}_{T^\star}) - f(\mathbf{u}_0) \leq f(\mathbf{u}_{T'}) - f(\mathbf{u}_0) \leq -2.5\mathscr{F}$$

Case $T' > T^\star$: In this case, by Lemma 16, we know $\|\mathbf{u}_t\| \leq O(\mathscr{S})$ for all $t \leq T^\star$. Define

$$T'' = \inf_t \left\{t | \tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) - f(\mathbf{w}_0) \leq -2\mathscr{F}\right\}$$

By Lemma 17, we immediately have $T'' \leq T^\star$. Apply same argument as in first case, we have for all $T \geq \frac{1}{c_{\max}}\mathscr{T}$ that $f(\mathbf{w}_T) - f(\mathbf{w}_0) \leq f(\mathbf{w}_{T^\star}) - f(\mathbf{w}_0) \leq -2.5\mathscr{F}$. □

Next we finish the proof by proving Lemma 16 and Lemma 17.

A.3.1. Proof of Lemma 16

In Lemma 16, we hope to show if the function value did not decrease, then all the iterations must be constrained in a small ball. We do that by analyzing the dynamics of the iterations, and we decompose the d-dimensional space into two subspaces: a subspace $\mathcal{S}$ which is the span of significantly negative eigenvectors of the Hessian and its orthogonal compliment.

Recall notation $\mathcal{H} := \nabla^2 f(\tilde{\mathbf{x}})$ and quadratic approximation $\tilde{f}_{\mathbf{y}}(\mathbf{x})$ as defined in Eq.(4). Since $\delta \in (0, \frac{d\kappa}{e}]$, we always have $\log(\frac{d\kappa}{\delta}) \geq 1$. W.L.O.G, set $\mathbf{u}_0 = 0$ to be the origin, by gradient descent update function, we have:

$$\begin{aligned}\mathbf{u}_{t+1} =&\mathbf{u}_t - \eta\nabla f(\mathbf{u}_t)\\ =&\mathbf{u}_t - \eta\nabla f(0) - \eta\left[\int_0^1 \nabla^2 f(\theta\mathbf{u}_t)\mathrm{d}\theta\right]\mathbf{u}_t\\ =&\mathbf{u}_t - \eta\nabla f(0) - \eta(\mathcal{H} + \Delta_t)\mathbf{u}_t\\ =&(\mathbf{I} - \eta\mathcal{H} - \eta\Delta_t)\mathbf{u}_t - \eta\nabla f(0) \end{aligned} \tag{5}$$

Here, $\Delta_t := \int_0^1 \nabla^2 f(\theta\mathbf{u}_t)\mathrm{d}\theta - \mathcal{H}$. By Hessian Lipschitz, we have $\|\Delta_t\| \leq \rho(\|\mathbf{u}_t\| + \|\tilde{\mathbf{x}}\|)$, and by smoothness of the gradient, we have $\|\nabla f(0)\| \leq \|\nabla f(\tilde{\mathbf{x}})\| + \ell\|\tilde{\mathbf{x}}\| \leq \mathscr{G} + \ell\cdot 2\mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta})) \leq 3\mathscr{G}$.

We will now compute the projections of $\mathbf{u}_t$ in different eigenspaces of $\mathcal{H}$. Let $\mathcal{S}$ be the subspace spanned by all eigenvectors of $\mathcal{H}$ whose eigenvalue is less than $-\frac{\gamma}{\hat{c}\log(\frac{d\kappa}{\delta})}$. $\mathcal{S}^c$ denotes the subspace of remaining eigenvectors. Let $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ denote the projections of $\mathbf{u}_t$ onto $\mathcal{S}$ and $\mathcal{S}^c$ respectively i.e., $\boldsymbol{\alpha}_t = \mathcal{P}_{\mathcal{S}}\mathbf{u}_t$, and $\boldsymbol{\beta}_t = \mathcal{P}_{\mathcal{S}^c}\mathbf{u}_t$. We can decompose the update equations Eq.(5) into:

$$\boldsymbol{\alpha}_{t+1} =(\mathbf{I} - \eta\mathcal{H})\boldsymbol{\alpha}_t - \eta\mathcal{P}_{\mathcal{S}}\Delta_t\mathbf{u}_t - \eta\mathcal{P}_{\mathcal{S}}\nabla f(0) \tag{6}$$

$$\boldsymbol{\beta}_{t+1} =(\mathbf{I} - \eta\mathcal{H})\boldsymbol{\beta}_t - \eta\mathcal{P}_{\mathcal{S}^c}\Delta_t\mathbf{u}_t - \eta\mathcal{P}_{\mathcal{S}^c}\nabla f(0) \tag{7}$$

By definition of T, we know for all $t < T$:

$$-3\mathscr{F} < \tilde{f}_0(\mathbf{u}_t) - f(0) =\nabla f(0)^\top\mathbf{u}_t - \frac{1}{2}\mathbf{u}_t^\top\mathcal{H}\mathbf{u}_t \leq \nabla f(0)^\top\mathbf{u}_t - \frac{\gamma}{2}\frac{\|\boldsymbol{\alpha}_t\|^2}{\hat{c}\log(\frac{d\kappa}{\delta})} + \frac{1}{2}\boldsymbol{\beta}_t^\top\mathcal{H}\boldsymbol{\beta}_t$$

Combined with the fact $\|\mathbf{u}_t\|^2 = \|\boldsymbol{\alpha}_t\|^2 + \|\boldsymbol{\beta}_t\|^2$, we have:

$$\begin{aligned}\|\mathbf{u}_t\|^2 \leq&\frac{2\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}\left(3\mathscr{F} + \nabla f(0)^\top\mathbf{u}_t + \frac{1}{2}\boldsymbol{\beta}_t^\top\mathcal{H}\boldsymbol{\beta}_t\right) + \|\boldsymbol{\beta}_t\|^2\\ \leq&14\cdot\max\left\{\frac{\mathscr{G}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}\|\mathbf{u}_t\|,\ \frac{\mathscr{F}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma},\ \frac{\boldsymbol{\beta}_t^\top\mathcal{H}\boldsymbol{\beta}_t\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma},\ \|\boldsymbol{\beta}_t\|^2\right\}\end{aligned}$$

where last inequality is due to $\|\nabla f(0)\| \leq 3\mathscr{G}$. This gives:

$$\|\mathbf{u}_t\| \leq 14\cdot\max\left\{\frac{\mathscr{G}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma},\ \sqrt{\frac{\mathscr{F}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}},\ \sqrt{\frac{\boldsymbol{\beta}_t^\top\mathcal{H}\boldsymbol{\beta}_t\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}},\ \|\boldsymbol{\beta}_t\|\right\} \tag{8}$$

Now, we use induction to prove that

$$\|\mathbf{u}_t\| \leq 100(\mathscr{S}\cdot\hat{c}) \tag{9}$$

Clearly Eq.(9) is true for $t = 0$ since $\mathbf{u}_0 = 0$. Suppose Eq.(9) is true for all $\tau \leq t$. We will now show that Eq.(9) holds for $t + 1 < T$. Note that by the definition of $\mathscr{S}$, $\mathscr{F}$ and $\mathscr{G}$, we only need to bound the last two terms of Eq.(8) i.e., $\|\boldsymbol{\beta}_{t+1}\|$ and $\boldsymbol{\beta}_{t+1}^\top\mathcal{H}\boldsymbol{\beta}_{t+1}$.

By update function of $\boldsymbol{\beta}_t$ (Eq.(7)), we have:

$$\boldsymbol{\beta}_{t+1} \leq(\mathbf{I} - \eta\mathcal{H})\boldsymbol{\beta}_t + \eta\boldsymbol{\delta}_t \tag{10}$$

and the norm of $\boldsymbol{\delta}_t$ is bounded as follows:

$$\begin{aligned}
\|\boldsymbol{\delta}_t\| &\leq \|\Delta_t\|\|\mathbf{u}_t\| + \|\nabla f(0)\| \\
&\leq \rho\left(\|\mathbf{u}_t\| + \|\tilde{\mathbf{x}}\|\right)\|\mathbf{u}_t\| + \|\nabla f(0)\| \\
&\leq \rho \cdot 100\hat{c}(100\hat{c} + 2/(\kappa \cdot \log(\frac{d\kappa}{\delta})))\mathscr{S}^2 + \mathscr{G} \\
&= [100\hat{c}(100\hat{c}+2)\sqrt{\eta\ell} + 1]\mathscr{G} \leq 2\mathscr{G}
\end{aligned} \tag{11}$$

The last step follows by choosing small enough constant $c_{\max} \leq \frac{1}{100\hat{c}(100\hat{c}+2)}$ and stepsize $\eta < c_{\max}/\ell$.

Bounding $\|\boldsymbol{\beta}_{t+1}\|$: Combining Eq.(10), Eq.(11) and using the definiton of $\mathcal{S}^c$, we have:

$$\|\boldsymbol{\beta}_{t+1}\| \leq (1 + \frac{\eta\gamma}{\hat{c}\log(\frac{d\kappa}{\delta})})\|\boldsymbol{\beta}_t\| + 2\eta\mathscr{G}$$

Since $\|\boldsymbol{\beta}_0\| = 0$ and $t + 1 \leq T$, by applying above relation recurrsively, we have:

$$\|\boldsymbol{\beta}_{t+1}\| \leq \sum_{\tau=0}^{t} 2(1 + \frac{\eta\gamma}{\hat{c}\log(\frac{d\kappa}{\delta})})^\tau \eta\mathscr{G} \leq 2 \cdot 3 \cdot T\eta\mathscr{G} \leq 6(\mathscr{S} \cdot \hat{c}) \tag{12}$$

The second last inequality is because $T \leq \hat{c}\mathscr{T}$ by definition, so that $(1 + \frac{\eta\gamma}{\hat{c}\log(\frac{d\kappa}{\delta})})^T \leq 3$.

Bounding $\boldsymbol{\beta}_{t+1}^\top \mathcal{H}\boldsymbol{\beta}_{t+1}$: Using Eq.(10), we can also write the update equation as:

$$\boldsymbol{\beta}_t = \sum_{\tau=0}^{t-1}(\mathbf{I} - \eta\mathcal{H})^\tau \boldsymbol{\delta}_\tau$$

Combining with Eq.(11), this gives

$$\begin{aligned}
\boldsymbol{\beta}_{t+1}^\top \mathcal{H}\boldsymbol{\beta}_{t+1} =& \eta^2 \sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t} \boldsymbol{\delta}_{\tau_1}^\top (\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}\boldsymbol{\delta}_{\tau_2} \\
\leq& \eta^2 \sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t} \|\boldsymbol{\delta}_{\tau_1}\|\|(\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}\|\|\boldsymbol{\delta}_{\tau_2}\| \\
\leq& 4\eta^2\mathscr{G}^2 \sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t} \|(\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}\|
\end{aligned}$$

Let the eigenvalues of $\mathcal{H}$ to be $\{\lambda_i\}$, then for any $\tau_1, \tau_2 \geq 0$, we know the eigenvalues of $(\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}$ are $\{\lambda_i(1 - \eta\lambda_i)^{\tau_1+\tau_2}\}$. Let $g_t(\lambda) := \lambda(1 - \eta\lambda)^t$, and setting its derivative to zero, we obtain:

$$\nabla g_t(\lambda) = (1 - \eta\lambda)^t - t\eta\lambda(1 - \eta\lambda)^{t-1} = 0$$

We see that $\lambda_t^\star = \frac{1}{(1+t)\eta}$ is the unique maximizer, and $g_t(\lambda)$ is monotonically increasing in $(-\infty, \lambda_t^\star]$. This gives:

$$\|(\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}\| = \max_i \lambda_i(1 - \eta\lambda_i)^{\tau_1+\tau_2} \leq \hat{\lambda}(1 - \eta\hat{\lambda})^{\tau_1+\tau_2} \leq \frac{1}{(1 + \tau_1 + \tau_2)\eta}$$

where $\hat{\lambda} = \min\{\ell, \lambda_{\tau_1+\tau_2}^\star\}$. Therefore, we have:

$$\begin{aligned}
\boldsymbol{\beta}_{t+1}^\top \mathcal{H}\boldsymbol{\beta}_{t+1} \leq& 4\eta^2\mathscr{G}^2 \sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t} \|(\mathbf{I} - \eta\mathcal{H})^{\tau_1}\mathcal{H}(\mathbf{I} - \eta\mathcal{H})^{\tau_2}\| \\
\leq& 4\eta\mathscr{G}^2 \sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t} \frac{1}{1 + \tau_1 + \tau_2} \leq 8\eta T\mathscr{G}^2 \leq 8\mathscr{S}^2\gamma\hat{c} \cdot \log^{-1}(\frac{d\kappa}{\delta})
\end{aligned} \tag{13}$$

The second last inequality is because by rearrange summation:

$$\sum_{\tau_1=0}^{t}\sum_{\tau_2=0}^{t}\frac{1}{1+\tau_1+\tau_2}=\sum_{\tau=0}^{2t}\min\{1+\tau,2t+1-\tau\}\cdot\frac{1}{1+\tau}\leq 2t+1<2T$$

Finally, substitue Eq.(12) and Eq.(13) into Eq.(8), this gives:

$$\begin{aligned}\|\mathbf{u}_{t+1}\|\leq&14\cdot\max\left\{\frac{\mathscr{G}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma},\ \sqrt{\frac{\mathscr{F}\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}},\ \sqrt{\frac{\boldsymbol{\beta}_t^\top\mathcal{H}\boldsymbol{\beta}_t\hat{c}\log(\frac{d\kappa}{\delta})}{\gamma}},\ \|\boldsymbol{\beta}_t\|\right\}\\ \leq&100(\mathscr{S}\cdot\hat{c})\end{aligned}$$

This finishes the induction as well as the proof of the lemma. □

A.3.2. Proof of Lemma 17

In this Lemma we try to show if all the iterates from $\mathbf{u}_0$ are constrained in a small ball, iterates from $\mathbf{w}_0$ must be able to decrease the function value. In order to do that, we keep track of vector $\mathbf{v}$ which is the difference between $\mathbf{u}$ and $\mathbf{w}$. Similar as before, we also decompose $\mathbf{v}$ into different eigenspaces. However, this time we only care about the projection of $\mathbf{v}$ on the direction $\mathbf{e}_1$ and its orthognal subspace.

Again, recall notation $\mathcal{H}:=\nabla^2 f(\tilde{\mathbf{x}})$, $\mathbf{e}_1$ as minimum eigenvector of $\mathcal{H}$ and quadratic approximation $\tilde{f}_{\mathbf{y}}(\mathbf{x})$ as defined in Eq.(4). Since $\delta\in(0,\frac{d\kappa}{e}]$, we always have $\log(\frac{d\kappa}{\delta})\geq 1$. W.L.O.G, set $\mathbf{u}_0=0$ to be the origin. Define $\mathbf{v}_t=\mathbf{w}_t-\mathbf{u}_t$, by assumptions in Lemma 17, we have $\mathbf{v}_0=\mu[\mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta}))]\mathbf{e}_1,\ \mu\in[\delta/(2\sqrt{d}),1]$. Now, consider the update equation for $\mathbf{w}_t$:

$$\begin{aligned}\mathbf{u}_{t+1}+\mathbf{v}_{t+1}=\mathbf{w}_{t+1}=&\mathbf{w}_t-\eta\nabla f(\mathbf{w}_t)\\ =&\mathbf{u}_t+\mathbf{v}_t-\eta\nabla f(\mathbf{u}_t+\mathbf{v}_t)\\ =&\mathbf{u}_t+\mathbf{v}_t-\eta\nabla f(\mathbf{u}_t)-\eta\left[\int_0^1\nabla^2 f(\mathbf{u}_t+\theta\mathbf{v}_t)\mathrm{d}\theta\right]\mathbf{v}_t\\ =&\mathbf{u}_t+\mathbf{v}_t-\eta\nabla f(\mathbf{u}_t)-\eta(\mathcal{H}+\Delta'_t)\mathbf{v}_t\\ =&\mathbf{u}_t-\eta\nabla f(\mathbf{u}_t)+(\mathbf{I}-\eta\mathcal{H}-\eta\Delta'_t)\mathbf{v}_t\end{aligned}$$

where $\Delta'_t:=\int_0^1\nabla^2 f(\mathbf{u}_t+\theta\mathbf{v}_t)\mathrm{d}\theta-\mathcal{H}$. By Hessian Lipschitz, we have $\|\Delta'_t\|\leq\rho(\|\mathbf{u}_t\|+\|\mathbf{v}_t\|+\|\tilde{\mathbf{x}}\|)$. This gives the dynamic for $\mathbf{v}_t$ satisfy:

$$\mathbf{v}_{t+1}=(\mathbf{I}-\eta\mathcal{H}-\eta\Delta'_t)\mathbf{v}_t \tag{14}$$

Since $\|\mathbf{w}_0-\tilde{\mathbf{x}}\|\leq\|\mathbf{u}_0-\tilde{\mathbf{x}}\|+\|\mathbf{v}_0\|\leq\mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta}))$, directly applying Lemma 16, we know $\mathbf{w}_t\leq 100(\mathscr{S}\cdot\hat{c})$ for all $t\leq T$. By condition of Lemma 17, we know $\|\mathbf{u}_t\|\leq 100(\mathscr{S}\cdot\hat{c})$ for all $t<T$. This gives:

$$\|\mathbf{v}_t\|\leq\|\mathbf{u}_t\|+\|\mathbf{w}_t\|\leq 200(\mathscr{S}\cdot\hat{c})\text{ for all }t<T \tag{15}$$

This in sum gives for $t<T$:

$$\|\Delta'_t\|\leq\rho(\|\mathbf{u}_t\|+\|\mathbf{v}_t\|+\|\tilde{\mathbf{x}}\|)\leq\rho(300\hat{c}\mathscr{S}+\mathscr{S}/(\kappa\cdot\log(\frac{d\kappa}{\delta})))\leq\rho\mathscr{S}(300\hat{c}+1)$$

On the other hand, denote ψ_t be the norm of $\mathbf{v}_t$ projected onto $\mathbf{e}_1$ direction, and φ_t be the norm of $\mathbf{v}_t$ projected onto remaining subspace. Eq.(14) gives us:

$$\begin{aligned}\psi_{t+1}\geq&(1+\gamma\eta)\psi_t-\mu\sqrt{\psi_t^2+\varphi_t^2}\\ \varphi_{t+1}\leq&(1+\gamma\eta)\varphi_t+\mu\sqrt{\psi_t^2+\varphi_t^2}\end{aligned}$$

where $\mu = \eta\rho\mathscr{S}(300\hat{c}+1)$. We will now prove via induction that for all $t < T$:

$$\varphi_t \leq 4\mu t \cdot \psi_t \tag{16}$$

By hypothesis of Lemma 17, we know $\varphi_0 = 0$, thus the base case of induction holds. Assume Eq.(16) is true for $\tau \leq t$, For $t+1 \leq T$, we have:

$$\begin{aligned} 4\mu(t+1)\psi_{t+1} \geq& 4\mu(t+1)\left((1+\gamma\eta)\psi_t - \mu\sqrt{\psi_t^2+\varphi_t^2}\right) \\ \varphi_{t+1} \leq& 4\mu t(1+\gamma\eta)\psi_t + \mu\sqrt{\psi_t^2+\varphi_t^2} \end{aligned}$$

From above inequalities, we see that we only need to show:

$$(1+4\mu(t+1))\sqrt{\psi_t^2+\varphi_t^2} \leq 4(1+\gamma\eta)\psi_t$$

By choosing $\sqrt{c_{\max}} \leq \frac{1}{300\hat{c}+1}\min\{\frac{1}{2\sqrt{2}}, \frac{1}{4\hat{c}}\}$, and $\eta \leq c_{\max}/\ell$, we have

$$4\mu(t+1) \leq 4\mu T \leq 4\eta\rho\mathscr{S}(300\hat{c}+1)\hat{c}\mathscr{T} = 4\sqrt{\eta\ell}(300\hat{c}+1)\hat{c} \leq 1$$

This gives:

$$4(1+\gamma\eta)\psi_t \geq 4\psi_t \leq 2\sqrt{2\psi_t^2} \geq (1+4\mu(t+1))\sqrt{\psi_t^2+\varphi_t^2}$$

which finishes the induction.

Now, we know $\varphi_t \leq 4\mu t \cdot \psi_t \leq \psi_t$, this gives:

$$\psi_{t+1} \geq (1+\gamma\eta)\psi_t - \sqrt{2}\mu\psi_t \geq (1+\frac{\gamma\eta}{2})\psi_t \tag{17}$$

where the last step follows from $\mu = \eta\rho\mathscr{S}(300\hat{c}+1) \leq \sqrt{c_{\max}}(300\hat{c}+1)\gamma\eta \cdot \log^{-1}(\frac{d\kappa}{\delta}) < \frac{\gamma\eta}{2\sqrt{2}}$.

Finally, combining Eq.(15) and (17) we have for all $t < T$:

$$\begin{aligned} 200(\mathscr{S}\cdot\hat{c}) \geq& \|\mathbf{v}_t\| \geq \psi_t \geq (1+\frac{\gamma\eta}{2})^t\psi_0 \\ =& (1+\frac{\gamma\eta}{2})^t c_0 \frac{\mathscr{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \geq (1+\frac{\gamma\eta}{2})^t \frac{\delta}{2\sqrt{d}}\frac{\mathscr{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \end{aligned}$$

This implies:

$$T < \frac{1}{2}\frac{\log(400\frac{\kappa\sqrt{d}}{\delta}\cdot\hat{c}\log(\frac{d\kappa}{\delta}))}{\log(1+\frac{\gamma\eta}{2})} \leq \frac{\log(400\frac{\kappa\sqrt{d}}{\delta}\cdot\hat{c}\log(\frac{d\kappa}{\delta}))}{\gamma\eta} \leq (2+\log(400\hat{c}))\mathscr{T}$$

The last inequality is due to $\delta \in (0, \frac{d\kappa}{e}]$ we have $\log(\frac{d\kappa}{\delta}) \geq 1$. By choosing constant $\hat{c}$ to be large enough to satisfy $2+\log(400\hat{c}) \leq \hat{c}$, we will have $T < \hat{c}\mathscr{T}$, which finishes the proof. □

B. Improve Convergence by Local Structure

In this section, we show if the objective function has nice *local structure* (e.g. satisfies Assumptions A3.a or A3.b), then it is possible to combine our analysis with the local analysis in order to get very fast convergence to a local minimum.

In particular, we prove Theorem 5.

Theorem 5. *There exist absolute constant $c_{\max}$ such that: if $f(\cdot)$ satisfies A1, A2, and A3.a (or A3.b), then for any $\delta > 0, \epsilon > 0, \Delta_f \geq f(\mathbf{x}_0) - f^\star$, constant $c \leq c_{\max}$, let $\tilde{\epsilon} = \min(\theta, \gamma^2/\rho)$, with probability $1-\delta$, the output of PGDli$(\mathbf{x}_0, \ell, \rho, \tilde{\epsilon}, c, \delta, \Delta_f, \beta)$ will be ϵ-close to $\mathcal{X}^\star$ in iterations:*

$$O\left(\frac{\ell(f(\mathbf{x}_0)-f^\star)}{\tilde{\epsilon}^2}\log^4\left(\frac{d\ell\Delta_f}{\tilde{\epsilon}^2\delta}\right) + \frac{\beta}{\alpha}\log\frac{\zeta}{\epsilon}\right)$$

Proof. Theorem 5 runs PGDli$(\mathbf{x}_0, \ell, \rho, \tilde{\epsilon}, c, \delta, \Delta_f, \beta)$. According to algorithm 3, we know it calls PGD$(\mathbf{x}_0, \ell, \rho, \epsilon, c, \delta, \Delta_f)$ first (denote its output as $\hat{\mathbf{x}}$), then run standard gradient descent with step size $\frac{1}{\beta}$ starting from $\hat{\mathbf{x}}$.

By Corollary 4, we know $\hat{\mathbf{x}}$ is already in the ζ-neighborhood of $\mathcal{X}^\star$, where $\mathcal{X}^\star$ is the set of local minima. Therefore, to prove this theorem, we only need to show prove following two claims:

1. Suppose $\{\mathbf{x}_t\}$ is the sequence of gradient descent starting from $\mathbf{x}_0 = \hat{\mathbf{x}}$ with step size $\frac{1}{\beta}$, then $\mathbf{x}_t$ is always in the ζ-neighborhood of $\mathcal{X}^\star$.

2. Local structure (assumption A3.a or A3.b) allows iterates to converge to points ϵ-close to $\mathcal{X}^\star$ within $O(\frac{\beta}{\alpha}\log\frac{\zeta}{\epsilon})$ iterations.

We will focus on Assumption A3.b (as we will later see Assumption A3.a is a special case of Assumption A3.b). Assume $\mathbf{x}_t$ is in ζ-neighborhood of $\mathcal{X}^\star$, by gradient updates and the definition of projection, we have:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_{t+1})\|^2 \leq & \|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2 = \|\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t) - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2 \\ = & \|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2 - 2\eta\langle\nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\rangle + \eta^2\|\nabla f(\mathbf{x}_t)\|^2 \\ \leq & \|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2 - \eta\alpha\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2 + (\eta^2 - \frac{\eta}{\beta})\|\nabla f(\mathbf{x})\|^2 \\ \leq & (1 - \frac{\alpha}{\beta})\|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2\end{aligned}$$

The second last inequality is due to (α, β)-regularity condition. The last inequality is because of the choice $\eta = \frac{1}{\beta}$.

There are two consequences of this calculation. First, it shows $\|\mathbf{x}_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_{t+1})\|^2 \leq \|\mathbf{x}_t - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_t)\|^2$. As a result if $\mathbf{x}_t$ in ζ-neighborhood of $\mathcal{X}^\star$, $\mathbf{x}_{t+1}$ is also in this ζ-neighborhood. Since $\mathbf{x}_0$ is in the ζ-neighborhood by Corollary 4, by induction we know all later iterations are in the same neighborhood.

Now, since we know all the points $\mathbf{x}_t$ are in the neighborhood, the equation also shows linear convergence rate $(1 - \frac{\alpha}{\beta})$. The initial distance is bounded by $\|\mathbf{x}_0 - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}_0)\| \leq \zeta$, therefore to converge to points ϵ-close to $\mathcal{X}^\star$, we only need the following number of iterations:

$$\frac{\log(\epsilon/\zeta)}{\log(1 - \alpha/\beta)} = O(\frac{\beta}{\alpha}\log\frac{\zeta}{\epsilon}).$$

This finishes the proof under Assumption A3.b.

Finally, we argue assumption A3.a implies A3.b. First, notice that if a function is locally strongly convex, then its local minima are isolated: for any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^\star$, the local region $B_{\mathbf{x}}(\zeta)$ and $B_{\mathbf{x}'}(\zeta)$ must be disjoint (otherwise function $f(\mathbf{x})$ is strongly convex in connected domain $B_{\mathbf{x}}(\zeta) \cup B_{\mathbf{x}'}(\zeta)$ but has two distinct local minima, which is impossible). Therefore, W.L.O.G, it suffices to consider one perticular disjoint region, with unique local minimum we denote as $\mathbf{x}^\star$, clearly, for all $\mathbf{x} \in B_{\mathbf{x}^\star}(\zeta)$ we have $\mathcal{P}_{\mathcal{X}^\star}(\mathbf{x}) = \mathbf{x}^\star$.

Now by α-strong convexity:

$$f(\mathbf{x}^\star) \geq f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2 \tag{18}$$

On the other hand, for any $\mathbf{x}$ in this ζ-neighborhood, we already proved $\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})$ also in this ζ-neighborhood. By β-smoothness, we also have:

$$f(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \frac{1}{2\beta}\|\nabla f(\mathbf{x})\|^2 \tag{19}$$

Combining Eq.(18) and Eq.(19), and using the fact $f(\mathbf{x}^\star) \leq f(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x}))$, we get:

$$\langle\nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2 + \frac{1}{2\beta}\|\nabla f(\mathbf{x})\|^2$$

which finishes the proof. □

C. Geometric Structures of Matrix Factorization Problem

In this Section we investigate the global geometric structures of the matrix factorization problem. These properties are summarized in Lemmas 6 and 7. Such structures allow us to apply our main Theorem and get fast convergence (as shown in Theorem 8).

Note that our main results Theorems 3 and 5 are proved for functions $f(\cdot)$ whose input $\mathbf{x}$ is a vector. For the current function in 2, though the input $\mathbf{U} \in \mathbb{R}^{d\times r}$ is a matrix, we can always vectorize it to be a vector in $\mathbb{R}^{dr}$ and apply our results. However, for simplicity of presentation, we still write everything in matrix form (without explicit vectorization), while the reader should keep in mind the operations are same if one vectorizes everything first.

Recall for vectors we use $\|\cdot\|$ to denote the 2-norm, and for matrices we use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ to denote spectral norm, and Frobenius norm respectively. Furthermore, we always use $\sigma_i(\cdot)$ to denote the i-th largest singular value of the matrix.

We first show how the geometric properties (Lemma 6 and Lemma 7) imply a fast convergence (Theorem 8).

Theorem 8. *There exists an absolute constant $c_{\max}$ such that the following holds. For matrix factorization (2), for any $\delta > 0$ and constant $c \le c_{\max}$, let $\Gamma^{1/2} := 2\max\{\|\mathbf{U}_0\|, 3(\sigma_1^\star)^{1/2}\}$, suppose we run PGDli$(\mathbf{U}_0, 8\Gamma, 12\Gamma^{1/2}, \frac{(\sigma_r^\star)^2}{108\Gamma^{1/2}}, c, \delta, \frac{r\Gamma^2}{2}, 10\sigma_1^\star)$, then:*

1. *With probability 1, the iterates satisfy $\|\mathbf{U}_t\| \le \Gamma^{1/2}$ for every $t \ge 0$.*
2. *With probability $1-\delta$, the output will be ϵ-close to global minima set $\mathcal{X}^\star$ in*

$$O\left(r\left(\frac{\Gamma}{\sigma_r^\star}\right)^4 \log^4\left(\frac{d\Gamma}{\delta\sigma_r^\star}\right) + \frac{\sigma_1^\star}{\sigma_r^\star}\log\frac{\sigma_r^\star}{\epsilon}\right)$$

iterations.

Proof of Theorem 8. Denote $\tilde{c}_{\max}$ to be the absolute constant allowed in Theorem 5. In this theorem, we let $c_{\max} = \min\{\tilde{c}_{\max}, \frac{1}{2}\}$, and choose any constant $c \le c_{\max}$.

Theorem 8 consists of two parts. In part 1 we show that the iterations never bring the matrix to a very large norm, while in part 2 we apply our main Theorem to get fast convergence. We will first prove the bound on number of iterations assuming the bound on the norm. We will then proceed to prove part 1.

Part 2: Assume part 1 of the theorem is true i.e., with probability 1, the iterates satisfy $\|\mathbf{U}_t\| \le \Gamma^{1/2}$ for every $t \ge 0$. In this case, although we are doing unconstrained optimization, we can still use the geometric properties that hold inside this region. .

By Lemma 6 and 7, we know objective function Eq.(2) is 8Γ-smooth, $12\Gamma^{1/2}$-Lipschitz Hessian, $(\frac{1}{24}(\sigma_r^\star)^{3/2}, \frac{1}{3}\sigma_r^\star, \frac{1}{3}(\sigma_r^\star)^{1/2})$-strict saddle, and holds $(\frac{2}{3}\sigma_r^\star, 10\sigma_1^\star)$-regularity condition in $\frac{1}{3}(\sigma_r^\star)^{1/2}$ neighborhood of local minima (also global minima) $\mathcal{X}^\star$. Furthermore, note $f^\star = 0$ and recall $\Gamma^{1/2} = 2\max\{\|\mathbf{U}_0\|, 3(\sigma_1^\star)^{1/2}\}$, then, we have:

$$f(\mathbf{U}_0) - f^\star = \|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{M}^\star\|_{\mathrm{F}}^2 \le 2r\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{M}^\star\|^2 \le \frac{r\Gamma^2}{2}.$$

Thus, we can choose $\Delta_f = \frac{r\Gamma^2}{2}$. Substituting the corresponding parameters from Theorem 5, we know by running PGDli$(\mathbf{U}_0, 8\Gamma, 12\Gamma^{1/2}, \frac{(\sigma_r^\star)^2}{108\Gamma^{1/2}}, c, \delta, \frac{r\Gamma^2}{2}, 10\sigma_1^\star)$, with probability $1-\delta$, the output will be ϵ-close to global minima set $\mathcal{X}^\star$ in iterations:

$$O\left(r\left(\frac{\Gamma}{\sigma_r^\star}\right)^4 \log^4\left(\frac{d\Gamma}{\delta\sigma_r^\star}\right) + \frac{\sigma_1^\star}{\sigma_r^\star}\log\frac{\sigma_r^\star}{\epsilon}\right).$$

Part 1: We will now show part 1 of the theorem. Recall PGDli (Algorithm 3) runs PGD (Algorithm 2) first, and then runs gradient descent within $\frac{1}{3}(\sigma_r^\star)^{1/2}$ neighborhood of $\mathcal{X}^\star$. It is easy to verify that $\frac{1}{3}(\sigma_r^\star)^{1/2}$ neighborhood of $\mathcal{X}^\star$ is a subset of $\{\mathbf{U}|\|\mathbf{U}\|^2 \le \Gamma\}$. Therefore, we only need to show that first phase PGD will not leave the region. Specifically, we now use induction to prove the following for PGD:

1. Suppose at iteration τ we add perturbation, and denote $\tilde{\mathbf{U}}_\tau$ to be the iterate before adding perturbation (i.e., $\mathbf{U}_\tau = \tilde{\mathbf{U}}_\tau + \xi_\tau$, and $\tilde{\mathbf{U}}_\tau = \mathbf{U}_{\tau-1} - \eta\nabla f(\mathbf{U}_{\tau-1})$). Then, $\|\tilde{\mathbf{U}}_\tau\| \le \frac{1}{2}\Gamma$, and

2. $\|\mathbf{U}_t\| \le \Gamma$ for all $t \ge 0$.

By choice of parameters of Algorithm 2, we know $\eta = \frac{c}{8\Gamma}$. First, consider gradient descent step without perturbations:

$$\begin{aligned}\|\mathbf{U}_{t+1}\| =&\|\mathbf{U}_t - \eta\nabla f(\mathbf{U}_t)\| = \|\mathbf{U}_t - \eta(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{M}^\star)\mathbf{U}_t\| \\ \le&\|\mathbf{U}_t - \eta\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_t\| + \eta\|\mathbf{M}^\star\mathbf{U}_t\| \\ \le&\max_i[\sigma_i(\mathbf{U}_t) - \eta\sigma_i^3(\mathbf{U}_t)] + \eta\|\mathbf{M}^\star\mathbf{U}_t\|\end{aligned}$$

For the first term, we know function $f(t) = t - \eta t^3$ is monotonically increasing in $[0, 1/\sqrt{3\eta}]$. On the other hand, by induction assumption, we also know $\|\mathbf{U}_t\| \le \Gamma^{1/2} \le \sqrt{8\Gamma/(3c)} = 1/\sqrt{3\eta}$. Therefore, the max is taken when $i = 1$:

$$\begin{aligned}\|\mathbf{U}_{t+1}\| \le&\|\mathbf{U}_t\| - \eta\|\mathbf{U}_t\|^3 + \eta\|\mathbf{M}^\star\mathbf{U}_t\| \\ \le&\|\mathbf{U}_t\| - \eta(\|\mathbf{U}_t\|^2 - \sigma_1^\star)\|\mathbf{U}_t\|. \end{aligned} \tag{20}$$

We seperate our discussion into following cases.

Case $\|\mathbf{U}_t\| > \frac{1}{2}\Gamma^{1/2}$: In this case $\|\mathbf{U}_t\| \ge \max\{\|\mathbf{U}_0\|, 3(\sigma_1^\star)^{1/2}\}$. Recall $\Gamma^{1/2} = 2\max\{\|\mathbf{U}_0\|, 3(\sigma_1^\star)^{1/2}\}$. Clearly, $\Gamma \ge 36\sigma_1^\star$, we know:

$$\begin{aligned}\|\mathbf{U}_{t+1}\| \le&\|\mathbf{U}_t\| - \eta(\|\mathbf{U}_t\|^2 - \sigma_1^\star)\|\mathbf{U}_t\| \le \|\mathbf{U}_t\| - \frac{c}{8\Gamma}(\frac{1}{4}\Gamma - \sigma_1^\star)\frac{1}{2}\Gamma^{1/2} \\ \le&\|\mathbf{U}_t\| - \frac{c}{8\Gamma}(\frac{1}{4}\Gamma - \frac{1}{36}\Gamma)\frac{1}{2}\Gamma^{1/2} = \|\mathbf{U}_t\| - \frac{c}{72}\Gamma^{1/2}.\end{aligned}$$

This means that in each iteration, the spectral norm would decrease by at least $\frac{c}{72}\Gamma^{1/2}$.

Case $\|\mathbf{U}_t\| \le \frac{1}{2}\Gamma^{1/2}$: From (20), we know that as long as $\|\mathbf{U}_t\|^2 \ge \|\mathbf{M}^\star\|$, we will always have $\|\mathbf{U}_{t+1}\| \le \|\mathbf{U}_t\| \le \frac{1}{2}\Gamma^{1/2}$. For $\|\mathbf{U}_t\|^2 \le \|\mathbf{M}^\star\|$, we have:

$$\begin{aligned}\|\mathbf{U}_{t+1}\| \le&\|\mathbf{U}_t\| - \eta(\|\mathbf{U}_t\|^2 - \sigma_1^\star)\|\mathbf{U}_t\| = \|\mathbf{U}_t\| + \frac{c}{8\Gamma}(\sigma_1^\star - \|\mathbf{U}_t\|^2)\|\mathbf{U}_t\| \\ \le&\|\mathbf{U}_t\| + ((\sigma_1^\star)^{1/2} - \|\mathbf{U}_t\|) \times \frac{c}{8\Gamma}((\sigma_1^\star)^{1/2} + \|\mathbf{U}_t\|)\|\mathbf{U}_t\| \\ \le&\|\mathbf{U}_t\| + ((\sigma_1^\star)^{1/2} - \|\mathbf{U}_t\|) \times \frac{c\sigma_1^\star}{4\Gamma} \le (\sigma_1^\star)^{1/2}\end{aligned}$$

Thus, in this case, we always have $\|\mathbf{U}_{t+1}\| \le \frac{1}{2}\Gamma^{1/2}$.

In conclusion, if we don't add perturbation in iteration t, we have:

- If $\|\mathbf{U}_t\| > \frac{1}{2}\Gamma^{1/2}$, then $\|\mathbf{U}_{t+1}\| \le \|\mathbf{U}_t\| - \frac{c}{72}\Gamma^{1/2}$.
- If $\|\mathbf{U}_t\| \le \frac{1}{2}\Gamma^{1/2}$, then $\|\mathbf{U}_{t+1}\| \le \frac{1}{2}\Gamma^{1/2}$.

Now consider the iterations where we add perturbation. By the choice of radius of perturbation in Algorithm 2 , we increase spectral norm by at most :

$$\|\xi_t\| \le \|\xi_t\|_{\mathrm{F}} \le \frac{\sqrt{c}}{\chi^2}\frac{(\sigma_r^\star)^2}{108\Gamma^{1/2}\cdot 8\Gamma} \le \frac{1}{2}\Gamma^{1/2}$$

The first inequality is because $\chi \ge 1$ and $c \le 1$. That is, if before perturbation we have $\|\tilde{\mathbf{U}}_t\| \le \frac{1}{2}\Gamma^{1/2}$, then $\|\mathbf{U}_t\| = \|\tilde{\mathbf{U}}_t + \xi_t\| \le \Gamma^{1/2}$.

On the other hand, according to Algorithm 2, once we add perturbation, we will not add perturbation for next $t_{\text{thres}} = \frac{\chi \cdot 24\Gamma}{c^2 \sigma_r^\star} \geq \frac{24}{c^2} \geq \frac{48}{c}$ iterations. Let $T = \min\{\inf_i\{\mathbf{U}_{t+i} | \|\mathbf{U}_{t+i}\| \leq \frac{1}{2}\Gamma^{1/2}\}, t_{\text{thres}}\}$:

$$\|\mathbf{U}_{t+T-1}\| \leq \|\mathbf{U}_t\| - \frac{c}{72}\Gamma^{1/2}(T-1) \leq \Gamma^{1/2}(1 - \frac{c(T-1)}{72})$$

This gives $T \leq \frac{36}{c} < \frac{48}{c} \leq t_{\text{thres}}$. Let $\tau > t$ be the next time when we add perturbation ($\tau \geq t + t_{\text{thres}}$), we immediately know $\|\mathbf{U}_{T+i}\| \leq \frac{1}{2}\Gamma^{1/2}$ for $0 \leq i < \tau - T$ and $\|\tilde{\mathbf{U}}_\tau\| \leq \frac{1}{2}\Gamma^{1/2}$.

Finally, $\|\mathbf{U}_0\| \leq \frac{1}{2}\Gamma^{1/2}$ by definition of Γ, so the initial condition holds. This finishes induction and the proof of the theorem. □

In the next subsections we prove the geometric structures.

C.1. Smoothness and Hessian Lipschitz

Before we start proofs of lemmas, we first state some properties about gradient and Hessians. The gradient of the objective function $f(\mathbf{U})$ is

$$\nabla f(\mathbf{U}) = 2(\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U}.$$

Furthermore, we have the gradient and Hessian satisfy for any $\mathbf{Z} \in \mathbb{R}^{d\times r}$:

$$\langle \nabla f(\mathbf{U}), \mathbf{Z}\rangle = 2\langle (\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U}, \mathbf{Z}\rangle, \text{ and} \tag{21}$$

$$\nabla^2 f(\mathbf{U})(\mathbf{Z}, \mathbf{Z}) = \|\mathbf{U}\mathbf{Z}^\top + \mathbf{Z}\mathbf{U}^\top\|_{\text{F}}^2 + 2\langle \mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star, \mathbf{Z}\mathbf{Z}^\top\rangle. \tag{22}$$

Lemma 6. *For any $\Gamma \geq \sigma_1^\star$, inside the region $\{\mathbf{U} | \|\mathbf{U}\|^2 < \Gamma\}$, $f(\mathbf{U})$ defined in Eq.(2) is 8Γ-smooth and $12\Gamma^{1/2}$-Hessian Lipschitz.*

Proof. Denote $\mathcal{D} = \{\mathbf{U} | \|\mathbf{U}\|^2 < \Gamma\}$, and recall $\Gamma \geq \sigma_1^\star$.

Part 1: For any $\mathbf{U}, \mathbf{V} \in \mathcal{D}$, we have:

$$\begin{aligned}\|\nabla f(\mathbf{U}) - \nabla f(\mathbf{V})\|_{\text{F}} =& 2\|(\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U} - (\mathbf{V}\mathbf{V}^\top - \mathbf{M}^\star)\mathbf{V}\|_{\text{F}} \\ \leq& 2\left[\|\mathbf{U}\mathbf{U}^\top\mathbf{U} - \mathbf{V}\mathbf{V}^\top\mathbf{V}\|_{\text{F}} + \|\mathbf{M}^\star(\mathbf{U} - \mathbf{V})\|_{\text{F}}\right] \\ \leq& 2\left[3\cdot\Gamma\|\mathbf{U} - \mathbf{V}\|_{\text{F}} + \sigma_1^\star\|\mathbf{U} - \mathbf{V}\|_{\text{F}}\right] \leq 8\Gamma \cdot \|\mathbf{U} - \mathbf{V}\|_{\text{F}}.\end{aligned}$$

The last line is due to following decomposition and triangle inequality:

$$\mathbf{U}\mathbf{U}^\top\mathbf{U} - \mathbf{V}\mathbf{V}^\top\mathbf{V} = \mathbf{U}\mathbf{U}^\top(\mathbf{U} - \mathbf{V}) + \mathbf{U}(\mathbf{U} - \mathbf{V})^\top\mathbf{V} + (\mathbf{U} - \mathbf{V})\mathbf{V}^\top\mathbf{V}.$$

Part 2: For any $\mathbf{U}, \mathbf{V} \in \mathcal{D}$, and any $\mathbf{Z} \in \mathbb{R}^{d\times r}$, according to Eq.(22), we have:

$$|\nabla^2 f(\mathbf{U})(\mathbf{Z}, \mathbf{Z}) - \nabla^2 f(\mathbf{V})(\mathbf{Z}, \mathbf{Z})| = \underbrace{\|\mathbf{U}\mathbf{Z}^\top + \mathbf{Z}\mathbf{U}^\top\|_{\text{F}}^2 - \|\mathbf{V}\mathbf{Z}^\top + \mathbf{Z}\mathbf{V}^\top\|_{\text{F}}^2}_{\mathfrak{A}} + \underbrace{2\langle \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top, \mathbf{Z}\mathbf{Z}^\top\rangle}_{\mathfrak{B}}.$$

For term $\mathfrak{A}$, we have:

$$\begin{aligned}\mathfrak{A} =& \langle \mathbf{U}\mathbf{Z}^\top + \mathbf{Z}\mathbf{U}^\top, (\mathbf{U} - \mathbf{V})\mathbf{Z}^\top + \mathbf{Z}(\mathbf{U} - \mathbf{V})^\top\rangle + \langle (\mathbf{U} - \mathbf{V})\mathbf{Z}^\top + \mathbf{Z}(\mathbf{U} - \mathbf{V})^\top, \mathbf{V}\mathbf{Z}^\top + \mathbf{Z}\mathbf{V}^\top\rangle \\ \leq& \|\mathbf{U}\mathbf{Z}^\top + \mathbf{Z}\mathbf{U}^\top\|_{\text{F}}\|(\mathbf{U} - \mathbf{V})\mathbf{Z}^\top + \mathbf{Z}(\mathbf{U} - \mathbf{V})^\top\|_{\text{F}} + \|(\mathbf{U} - \mathbf{V})\mathbf{Z}^\top + \mathbf{Z}(\mathbf{U} - \mathbf{V})^\top\|_{\text{F}}\|\mathbf{V}\mathbf{Z}^\top + \mathbf{Z}\mathbf{V}^\top\|_{\text{F}} \\ \leq& 4\|\mathbf{U}\|\|\mathbf{Z}\|_{\text{F}}^2\|\mathbf{U} - \mathbf{V}\|_{\text{F}} + 4\|\mathbf{V}\|\|\mathbf{Z}\|_{\text{F}}^2\|\mathbf{U} - \mathbf{V}\|_{\text{F}} \leq 8\Gamma^{1/2}\|\mathbf{Z}\|_{\text{F}}^2\|\mathbf{U} - \mathbf{V}\|_{\text{F}}.\end{aligned}$$

For term $\mathfrak{B}$, we have:

$$\mathfrak{B} \leq 2\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_\mathrm{F}\|\mathbf{Z}\mathbf{Z}^\top\|_\mathrm{F} \leq 4\Gamma^{1/2}\|\mathbf{Z}\|_\mathrm{F}^2\|\mathbf{U}-\mathbf{V}\|_\mathrm{F}.$$

The inequality is due to following decomposition and triangle inequality:

$$\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top = \mathbf{U}(\mathbf{U}-\mathbf{V})^\top + (\mathbf{U}-\mathbf{V})\mathbf{V}^\top.$$

Therefore, in sum we have:

$$\begin{aligned}\max_{\mathbf{Z}:\|\mathbf{Z}\|_\mathrm{F}\leq 1} |\nabla^2 f(\mathbf{U})(\mathbf{Z},\mathbf{Z}) - \nabla^2 f(\mathbf{V})(\mathbf{Z},\mathbf{Z})| \leq& \max_{\mathbf{Z}:\|\mathbf{Z}\|_\mathrm{F}\leq 1} 12\Gamma^{1/2}\|\mathbf{Z}\|_\mathrm{F}^2\|\mathbf{U}-\mathbf{V}\|_\mathrm{F} \\ \leq& 12\Gamma^{1/2}\|\mathbf{U}-\mathbf{V}\|_\mathrm{F}.\end{aligned}$$

□

C.2. Strict-Saddle Property and Local Regularity

Recall the gradient and Hessian of objective function is calculated as in Eq.(21) and Eq.(22). We first prove an elementary inequality regarding to the trace of product of two symmetric PSD matrices. This lemma will be frequently used in the proof.

Lemma 18. *For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ both symmetric PSD matrices, we have:*

$$\sigma_{\min}(\mathbf{A})tr(\mathbf{B}) \leq tr(\mathbf{A}\mathbf{B}) \leq \|\mathbf{A}\|tr(\mathbf{B})$$

Proof. Let $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ be the eigendecomposition of $\mathbf{A}$, where $\mathbf{D}$ is diagonal matrix, and $\mathbf{V}$ is orthogonal matrix. Then we have:

$$\mathrm{tr}(\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{D}\mathbf{V}^\top\mathbf{B}\mathbf{V}) = \sum_{i=1}^d \mathbf{D}_{ii}(\mathbf{V}^\top\mathbf{B}\mathbf{V})_{ii}.$$

Since $\mathbf{B}$ is PSD, we know $\mathbf{V}^\top\mathbf{B}\mathbf{V}$ is also PSD, thus the diagonal entries are non-negative. That is, $(\mathbf{V}^\top\mathbf{B}\mathbf{V})_{ii} \geq 0$ for all $i = 1, \ldots, d$. Finally, the lemma follows from the fact that $\sigma_{\min}(\mathbf{A}) \leq \mathbf{D}_{ii} \leq \|\mathbf{A}\|$ and $\mathrm{tr}(\mathbf{V}^\top\mathbf{B}\mathbf{V}) = \mathrm{tr}(\mathbf{B}\mathbf{V}\mathbf{V}^\top) = \mathrm{tr}(\mathbf{B})$. □

Now, we are ready to prove Lemma 7.

Lemma 7. *For $f(\mathbf{U})$ defined in Eq.(2), all local minima are global minima. The set of global minima is $\mathcal{X}^\star = \{\mathbf{V}^\star\mathbf{R}|\mathbf{R}\mathbf{R}^\top = \mathbf{R}^\top\mathbf{R} = \mathbf{I}\}$. Furthermore, $f(\mathbf{U})$ satisfies:*

1. *$(\frac{1}{24}(\sigma_r^\star)^{3/2}, \frac{1}{3}\sigma_r^\star, \frac{1}{3}(\sigma_r^\star)^{1/2})$-strict saddle property, and*
2. *$(\frac{2}{3}\sigma_r^\star, 10\sigma_1^\star)$-regularity condition in $\frac{1}{3}(\sigma_r^\star)^{1/2}$ neighborhood of $\mathcal{X}^\star$.*

Proof. Let us denote the set $\mathcal{X}^\star := \{\mathbf{V}^\star\mathbf{R}|\mathbf{R}\mathbf{R}^\top = \mathbf{R}^\top\mathbf{R} = \mathbf{I}\}$, in the end of proof, we will show this set is the set of all local minima (which is also global minima).

Throughout the proof of this lemma, we always focus on the first-order and second-order property for one matrix $\mathbf{U}$. For simplicity of calculation, when it is clear from the context, we denote $\mathbf{U}^\star = \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})$ and $\Delta = \mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})$. By definition of $\mathcal{X}^\star$, we know $\mathbf{U}^\star = \mathbf{V}^\star\mathbf{R}_\mathbf{U}$ and $\Delta = \mathbf{U} - \mathbf{V}^\star\mathbf{R}_\mathbf{U}$, where

$$\mathbf{R}_\mathbf{U} = \operatorname*{argmin}_{\mathbf{R}:\mathbf{R}\mathbf{R}^\top = \mathbf{R}^\top\mathbf{R} = \mathbf{I}} \|\mathbf{U} - \mathbf{V}^\star\mathbf{R}\|_\mathrm{F}^2$$

We first prove following claim, which will used in many places across this proof:

$$\mathbf{U}^\top\mathbf{U}^\star = \mathbf{U}^\top\mathbf{V}^\star\mathbf{R}_\mathbf{U} \text{ is a symmetric PSD matrix.} \tag{23}$$

This because by expanding the Frobenius norm, and letting the SVD of $\mathbf{V}^{\star\top}\mathbf{U}$ be $\mathbf{A}\mathbf{D}\mathbf{B}^\top$, we have:

$$\begin{aligned}\underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\text{argmin}} \|\mathbf{U}-\mathbf{V}^\star\mathbf{R}\|_\text{F}^2 =& \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\text{argmin}} -\langle\mathbf{U},\mathbf{V}^\star\mathbf{R}\rangle\\ = \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\text{argmin}} -\text{tr}(\mathbf{U}^\top\mathbf{V}^\star\mathbf{R}) =& \underset{\mathbf{R}:\mathbf{R}\mathbf{R}^\top=\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\text{argmin}} -\text{tr}(\mathbf{D}\mathbf{A}^\top\mathbf{R}\mathbf{B})\end{aligned}$$

Since $\mathbf{A},\mathbf{B},\mathbf{R}$ are all orthonormal matrix, we know $\mathbf{A}^\top\mathbf{R}\mathbf{B}$ is also orthonormal matrix. Moreover for any orthonormal matrix $\mathbf{T}$, we have:

$$\text{tr}(\mathbf{D}\mathbf{T}) = \sum_i \mathbf{D}_{ii}\mathbf{T}_{ii} \le \sum_i \mathbf{D}_{ii}$$

The last inequality is because $\mathbf{D}_{ii}$ is singular value thus non-negative, and $\mathbf{T}$ is orthonormal, thus $\mathbf{T}_{ii} \le 1$. This means the maximum of tr$(\mathbf{D}\mathbf{T})$ is achieved when $\mathbf{T} = \mathbf{I}$, i.e., the minimum of $-\text{tr}(\mathbf{D}\mathbf{A}^\top\mathbf{R}\mathbf{B})$ is achieved when $\mathbf{R} = \mathbf{A}\mathbf{B}^\top$. Therefore, $\mathbf{U}^\top\mathbf{V}^\star\mathbf{R}_\mathbf{U} = \mathbf{B}\mathbf{D}\mathbf{A}^\top\mathbf{A}\mathbf{B}^\top = \mathbf{B}\mathbf{D}\mathbf{B}^\top$ is symmetric PSD matrix.

Part 1: In order to show the strict saddle property, we only need to show that for any $\mathbf{U}$ satisfying $\|\nabla f(\mathbf{U})\|_\text{F} \le \frac{1}{24}(\sigma_r^\star)^{3/2}$ and $\|\Delta\|_\text{F} = \|\mathbf{U}-\mathbf{U}^\star\|_\text{F} \ge \frac{1}{3}(\sigma_r^\star)^{1/2}$, we always have $\sigma_{\min}(\nabla^2 f(\mathbf{U})) \le -\frac{1}{3}\sigma_r^\star$.

Let's consider Hessian $\nabla^2(\mathbf{U})$ in the direction of $\Delta = \mathbf{U}-\mathbf{U}^\star$. Clearly, we have:

$$\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star = \mathbf{U}\mathbf{U}^\top - (\mathbf{U}-\Delta)(\mathbf{U}-\Delta)^\top = (\mathbf{U}\Delta^\top + \Delta\mathbf{U}^\top) - \Delta\Delta^\top$$

and by (21):

$$\begin{aligned}\langle\nabla f(\mathbf{U}),\Delta\rangle =& 2\langle(\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star)\mathbf{U},\Delta\rangle = \langle\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star, \Delta\mathbf{U}^\top+\mathbf{U}\Delta^\top\rangle\\ =& \langle\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star, \mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star+\Delta\Delta^\top\rangle\end{aligned}$$

Therefore, by Eq.(22) and above two equalities, we have:

$$\begin{aligned}\nabla^2 f(\mathbf{U})(\Delta,\Delta) =& \|\mathbf{U}\Delta^\top+\Delta\mathbf{U}^\top\|_\text{F}^2 + 2\langle\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star,\Delta\Delta^\top\rangle\\ =& \|\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star+\Delta\Delta^\top\|_\text{F}^2 + 2\langle\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star,\Delta\Delta^\top\rangle\\ =& \|\Delta\Delta^\top\|_\text{F}^2 - 3\|\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star\|_\text{F}^2 + 4\langle\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star,\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star+\Delta\Delta^\top\rangle\\ =& \|\Delta\Delta^\top\|_\text{F}^2 - 3\|\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star\|_\text{F}^2 + 4\langle\nabla f(\mathbf{U}),\Delta\rangle\end{aligned}$$

Consider the first two terms, by expanding, we have:

$$\begin{aligned}&3\|\mathbf{U}\mathbf{U}^\top-\mathbf{M}^\star\|_\text{F}^2 - \|\Delta\Delta^\top\|_\text{F}^2 = 3\|(\mathbf{U}^\star\Delta^\top+\Delta\mathbf{U}^{\star\top})+\Delta\Delta^\top\|_\text{F}^2 - \|\Delta\Delta^\top\|_\text{F}^2\\ =&3\cdot\text{tr}\left(2\mathbf{U}^{\star\top}\mathbf{U}^\star\Delta^\top\Delta + 2(\mathbf{U}^{\star\top}\Delta)^2 + 4\mathbf{U}^{\star\top}\Delta\Delta^\top\Delta + (\Delta^\top\Delta)^2\right) - \text{tr}((\Delta^\top\Delta)^2)\\ =&\text{tr}\left(6\mathbf{U}^{\star\top}\mathbf{U}^\star\Delta^\top\Delta + 6(\mathbf{U}^{\star\top}\Delta)^2 + 12\mathbf{U}^{\star\top}\Delta\Delta^\top\Delta + 2(\Delta^\top\Delta)^2\right)\\ =&\text{tr}((4\sqrt{3}-6)\mathbf{U}^{\star\top}\mathbf{U}^\star\Delta^\top\Delta + (12-4\sqrt{3})\mathbf{U}^{\star\top}(\mathbf{U}^\star+\Delta)\Delta^\top\Delta + 2(\sqrt{3}\mathbf{U}^{\star\top}\Delta+\Delta^\top\Delta)^2)\\ \ge&(4\sqrt{3}-6)\text{tr}(\mathbf{U}^{\star\top}\mathbf{U}^\star\Delta^\top\Delta) \ge (4\sqrt{3}-6)\sigma_r^\star\|\Delta\|_\text{F}^2\end{aligned}$$

where the second last inequality is because $\mathbf{U}^{\star\top}(\mathbf{U}^\star+\Delta)\Delta^\top\Delta = \mathbf{U}^{\star\top}\mathbf{U}\Delta^\top\Delta$ is the product of two symmetric PSD matrices (thus its trace is non-negative); the last inequality is by Lemma 18.

Finally, in case we have $\|\nabla f(\mathbf{U})\|_\text{F} \le \frac{1}{24}(\sigma_r^\star)^{3/2}$ and $\|\Delta\|_\text{F} = \|\mathbf{U}-\mathbf{U}^\star\|_\text{F} \ge \frac{1}{3}(\sigma_r^\star)^{1/2}$

$$\begin{aligned}\sigma_{\min}(\nabla^2 f(\mathbf{U})) \le& \frac{1}{\|\Delta\|_\text{F}^2}\nabla^2 f(\mathbf{U})(\Delta,\Delta) \le -(4\sqrt{3}-6)\sigma_r^\star + 4\frac{\langle\nabla f(\mathbf{U}),\Delta\rangle}{\|\Delta\|_\text{F}^2}\\ \le& -(4\sqrt{3}-6)\sigma_r^\star + 4\frac{\|\nabla f(\mathbf{U})\|_\text{F}}{\|\Delta\|_\text{F}} \le -(4\sqrt{3}-6.5)\sigma_r^\star \le -\frac{1}{3}\sigma_r^\star\end{aligned}$$

Part 2: In $\frac{1}{3}(\sigma_r^\star)^{1/2}$ neigborhood of $\mathcal{X}^\star$, by definition, we know,

$$\|\Delta\|_{\text{F}}^2 = \|\mathbf{U} - \mathbf{U}^\star\|_{\text{F}}^2 \le \frac{1}{9}\sigma_r^\star.$$

Clearly, by Weyl's inequality, we have $\|\mathbf{U}\| \le \|\mathbf{U}^\star\| + \|\Delta\| \le \frac{4}{3}(\sigma_1^\star)^{1/2}$, and $\sigma_r(\mathbf{U}) \ge \sigma_r(\mathbf{U}^\star) - \|\Delta\| \ge \frac{2}{3}(\sigma_r^\star)^{1/2}$. Moreover, since $\mathbf{U}^{\star\top}\mathbf{U}$ is symmetric matrix, we have:

$$\begin{aligned}
\sigma_r(\mathbf{U}^{\star\top}\mathbf{U}) =& \frac{1}{2}\left(\sigma_r(\mathbf{U}^\top\mathbf{U}^\star + \mathbf{U}^{\star\top}\mathbf{U})\right) \\
\ge& \frac{1}{2}\left(\sigma_r(\mathbf{U}^\top\mathbf{U} + \mathbf{U}^{\star\top}\mathbf{U}^\star) - \|(\mathbf{U}-\mathbf{U}^\star)^\top(\mathbf{U}-\mathbf{U}^\star)\|\right) \\
\ge& \frac{1}{2}\left(\sigma_r(\mathbf{U}^\top\mathbf{U}) + \sigma_r(\mathbf{U}^{\star\top}\mathbf{U}^\star) - \|\Delta\|_{\text{F}}^2\right) \\
\ge& \frac{1}{2}(1 + \frac{4}{9} - \frac{1}{9})\sigma_r^\star = \frac{2}{3}\sigma_r^\star.
\end{aligned}$$

At a highlevel, we will prove (α, β)-regularity property (1) by proving that:

1. $\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x})\rangle \ge \alpha\|\mathbf{x} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x})\|^2$, and
2. $\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{x})\rangle \ge \frac{1}{\beta}\|\nabla f(\mathbf{x})\|^2$.

According to (21), we know:

$$\begin{aligned}
\langle \nabla f(\mathbf{U}), \mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})\rangle =& 2\langle(\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U}, \Delta\rangle = 2\langle \mathbf{U}\Delta^\top + \Delta\mathbf{U}^{\star\top}, \Delta\mathbf{U}^\top\rangle \\
=& 2(\text{tr}(\mathbf{U}\Delta^\top\mathbf{U}\Delta^\top) + \text{tr}(\Delta\mathbf{U}^{\star\top}\mathbf{U}\Delta^\top)) \\
=& 2(\|\Delta^\top\mathbf{U}\|_{\text{F}}^2 + \text{tr}(\mathbf{U}^{\star\top}\mathbf{U}\Delta^\top\Delta)).
\end{aligned} \tag{24}$$

The last equality is because $\Delta^\top\mathbf{U}$ is symmetric matrix. Since $\mathbf{U}^{\star\top}\mathbf{U}$ is symmetric PSD matrix, and recall $\sigma_r(\mathbf{U}^{\star\top}\mathbf{U}) \ge \frac{2}{3}\sigma_r^\star$, by Lemma 18 we have:

$$\langle \nabla f(\mathbf{U}), \mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})\rangle \ge \sigma_r(\mathbf{U}^{\star\top}\mathbf{U})\text{tr}(\Delta^\top\Delta) \ge \frac{2}{3}\sigma_r^\star\|\Delta\|_{\text{F}}^2. \tag{25}$$

On the other hand, we also have:

$$\begin{aligned}
\|\nabla f(\mathbf{U})\|_{\text{F}}^2 =& 4\langle(\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U}, (\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star)\mathbf{U}\rangle \\
=& 4\langle(\mathbf{U}\Delta^\top + \Delta\mathbf{U}^{\star\top})\mathbf{U}, (\mathbf{U}\Delta^\top + \Delta\mathbf{U}^{\star\top})\mathbf{U}\rangle \\
=& 4\left(\underbrace{\text{tr}[(\Delta^\top\mathbf{U}\mathbf{U}^\top\Delta)\mathbf{U}^\top\mathbf{U}] + 2\text{tr}[\Delta^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}^\star\Delta^\top\mathbf{U}]}_{\mathfrak{A}} + \underbrace{\text{tr}(\mathbf{U}^{\star\top}\mathbf{U}\mathbf{U}^\top\mathbf{U}^\star\Delta^\top\Delta)}_{\mathfrak{B}}\right).
\end{aligned}$$

For term $\mathfrak{A}$, by Lemma 18, and $\Delta^\top\mathbf{U}$ being a symmetric matrix, we have:

$$\mathfrak{A} \le \|\mathbf{U}^\top\mathbf{U}\|\|\Delta^\top\mathbf{U}\|_{\text{F}}^2 + 2\|\mathbf{U}^\top\mathbf{U}^\star\|\|\Delta^\top\mathbf{U}\|_{\text{F}}^2 \le (\frac{16}{9} + \frac{8}{3})\sigma_1^\star\|\Delta^\top\mathbf{U}\|_{\text{F}}^2 \le 5\sigma_1^\star\|\Delta^\top\mathbf{U}\|_{\text{F}}^2$$

For term $\mathfrak{B}$, by Eq.(23) we can denote $\mathbf{C} = \mathbf{U}^{\star\top}\mathbf{U} = \mathbf{U}^\top\mathbf{U}^\star$ which is symmetric PSD matrix, by Lemma 18, we have:

$$\begin{aligned}
\mathfrak{B} =& \text{tr}(\mathbf{C}^2\Delta^\top\Delta) = \text{tr}(\mathbf{C}(\mathbf{C}^{1/2}\Delta^\top\Delta\mathbf{C}^{1/2})) \\
\le& \|\mathbf{C}\|\text{tr}(\mathbf{C}^{1/2}\Delta^\top\Delta\mathbf{C}^{1/2}) = \|\mathbf{C}\|\text{tr}(\mathbf{C}\Delta^\top\Delta) \le \frac{4}{3}\sigma_1^\star\text{tr}(\mathbf{U}^{\star\top}\mathbf{U}\Delta^\top\Delta).
\end{aligned}$$

Combining with (24) we have:

$$\|\nabla f(\mathbf{U})\|_{\text{F}}^2 \le \sigma_1^\star(20\|\Delta^\top\mathbf{U}\|_{\text{F}}^2 + \frac{16}{3}\text{tr}(\mathbf{U}^{\star\top}\mathbf{U}\Delta^\top\Delta)) \le 10\sigma_1^\star\langle \nabla f(\mathbf{U}), \mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})\rangle. \tag{26}$$

Combining (25) and (26), we have:

$$\langle \nabla f(\mathbf{U}), \mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U}) \rangle \geq \frac{1}{3}\sigma_r^\star \|\mathbf{U} - \mathcal{P}_{\mathcal{X}^\star}(\mathbf{U})\|_{\mathrm{F}}^2 + \frac{1}{20\sigma_1^\star}\|\nabla f(\mathbf{U})\|_{\mathrm{F}}^2.$$

□