Noisy Gradient Descent Converges to Flat Minima for Nonconvex Matrix Factorization: Supplementary Materials

1 Preliminaries

We first introduce some important notions and results which can be used in the following proof.

Assuming $\{u_*, \widetilde{u}_1, \ldots, \widetilde{u}_{d_1-1}\}$ and $\{v_*, \widetilde{v}_1, \ldots, \widetilde{v}_{d_2-1}\}$ are two sets of standard orthogonal basis of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively, we can then rewrite $\forall x \in \mathbb{R}^{d_1}$ and $\forall y \in \mathbb{R}^{d_2}$ as

$$x \triangleq \alpha_1 u_* + \sum_{i=1}^{d_1-1} \beta_1^{(i)} \widetilde{u}_i,$$
$$y \triangleq \alpha_2 v_* + \sum_{j=1}^{d_2-1} \beta_2^{(j)} \widetilde{v}_j,$$

where $\alpha_1 = x^\top u_*$, $\alpha_2 = y^\top v_*$, $\beta_1^{(i)} = x^\top \widetilde{u}_i$ and $\beta_2^{(j)} = y^\top \widetilde{v}_j$, $\forall 0 \le i \le d_1 - 1, 0 \le j \le d_2 - 1$. For simplicity, we denote $\beta_k = (\beta_k^{(1)}, ..., \beta_k^{(d_k-1)})^\top$ where $k = 1, 2$.

With the notions above, we can rewrite the Perturbed GD update as

$$\alpha_{1,t+1} = \alpha_{1,t} - \eta < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >,$$
$$\alpha_{2,t+1} = \alpha_{2,t} - \eta < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* > .$$

Note that the optimal solutions to (2) satisfy $x^\top My = 1$. Thus, in our proof, we need to characterize the update of $x^\top My$, which can be re-expressed as

$$\begin{aligned} x_{t+1}^\top M y_{t+1} &= \alpha_{1,t+1}\alpha_{2,t+1} \\ &= (\alpha_{1,t} - \eta < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >)(\alpha_{2,t} - \eta < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >) \\ &= \alpha_{1,t}\alpha_{2,t} - \eta \left(\alpha_{2,t} < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* > + \alpha_{1,t} < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >\right) \\ &\quad + \eta^2 < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >< \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* > . \end{aligned} \tag{1}$$

For simplicity, we denote

$$A_t \triangleq \alpha_{2,t} < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* > + \alpha_{1,t} < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >,$$
$$B_t \triangleq < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >< \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* > .$$

Then the update of $x^\top My = \alpha_1\alpha_2$ can be expressed in a more compact way as follows.

$$\alpha_{1,t+1}\alpha_{2,t+1} = \alpha_{1,t}\alpha_{2,t} - \eta A_t + \eta^2 B_t. \tag{2}$$

Similarly, the update of $(x^\top My - 1)^2$ can be re-expressed as

$$\begin{aligned} (x_{t+1}^\top M y_{t+1} - 1)^2 &= (\alpha_{1,t+1}\alpha_{2,t+1} - 1)^2 \\ &= (\alpha_{1,t}\alpha_{2,t} - 1 - \eta A_t + \eta^2 B_t)^2 \\ &= (\alpha_{1,t}\alpha_{2,t} - 1)^2 + \eta^2 A_t^2 + \eta^4 B_t^2 \\ &\quad - 2\eta A_t(\alpha_{1,t}\alpha_{2,t} - 1) - 2\eta^3 A_t B_t + 2\eta^2 B_t(\alpha_{1,t}\alpha_{2,t} - 1). \end{aligned} \tag{3}$$

Furthermore, since the balanced optima satisfy $x^\top u_* = y^\top v_*$, we further explicitly write down the update of $\left((x^\top u_*)^2 - (y^\top v_*)^2\right)^2$ as follows.

$$\begin{aligned}\left((x_{t+1}^\top u_*)^2 - (y_{t+1}^\top v_*)^2\right)^2 &= (\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 \\ &= \left(\alpha_{1,t}^2 - \alpha_{2,t}^2\right)^2 + 4\eta^2 D_t^2 + \eta^4 F_t^2 \\ &\quad - 4\eta D_t\left(\alpha_{1,t}^2 - \alpha_{2,t}^2\right) - 4\eta^3 D_t F_t + 2\eta^2 F_t\left(\alpha_{1,t}^2 - \alpha_{2,t}^2\right), \end{aligned} \tag{4}$$

where D_t and F_t is defined as

$$\begin{aligned} D_t &\triangleq \alpha_{1,t} < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* > - \alpha_{2,t} < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >, \\ F_t &\triangleq < \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >^2 - < \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >^2 . \end{aligned}$$

Next we are going to calculate $< \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >$, the gradient projection along the direction of the optimum u_*.

$$\begin{aligned} &< \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* > \\ &= u_*^\top \nabla_x \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}) \\ &= u_*^\top \left((x_t + \xi_{1,t})(y_t + \xi_{2,t})^\top - u_* v_*^\top\right)(y_t + \xi_{2,t}) \\ &= \alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|_2^2) - \alpha_{2,t} + 2\alpha_{1,t} y_t^\top \xi_{2,t} - v_*^\top \xi_{2,t} + u_*^\top \xi_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|_2^2) \\ &\qquad + 2u_*^\top \xi_{1,t} y_t^\top \xi_{2,t} + \alpha_{1,t}\|\xi_{2,t}\|^2 + u_*^\top \xi_{1,t}\|\xi_{2,t}\|^2 \\ &\triangleq \alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|_2^2) - \alpha_{2,t} + g_x, \end{aligned} \tag{5}$$

where $g_x = 2\alpha_{1,t} y_t^\top \xi_{2,t} - v_*^\top \xi_{2,t} + u_*^\top \xi_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|_2^2) + 2u_*^\top \xi_{1,t} y_t^\top \xi_{2,t} + \alpha_{1,t}\|\xi_{2,t}\|^2 + u_*^\top \xi_{1,t}\|\xi_{2,t}\|^2$.

Similarly, we have

$$\begin{aligned} &< \nabla_y \mathcal{F}(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* > \\ &= \alpha_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|_2^2) - \alpha_{1,t} + 2\alpha_{2,t} x_t^\top \xi_{1,t} - u_*^\top \xi_{1,t} + v_*^\top \xi_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|_2^2) \\ &\qquad + 2v_*^\top \xi_{2,t} x_t^\top \xi_{1,t} + \alpha_{2,t}\|\xi_{1,t}\|^2 + v_*^\top \xi_{2,t}\|\xi_{1,t}\|^2 \\ &\triangleq \alpha_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|_2^2) - \alpha_{1,t} + g_y, \end{aligned} \tag{6}$$

where $g_y = \alpha_{1,t} + 2\alpha_{2,t} x_t^\top \xi_{1,t} - u_*^\top \xi_{1,t} + v_*^\top \xi_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|_2^2) + 2v_*^\top \xi_{2,t} x_t^\top \xi_{1,t} + \alpha_{2,t}\|\xi_{1,t}\|^2 + v_*^\top \xi_{2,t}\|\xi_{1,t}\|^2$.

2 Proof of Lemma 1, Lemma 2 , and Lemma 3

2.1 Proof of Lemma 1

Proof. By setting the gradient of $\mathcal{F}$ to zero we get

$$\|x\|_2^2 y = M^\top x, \tag{7}$$

$$\|y\|_2^2 x = My. \tag{8}$$

Recall that $M = u_* v_*^\top$ and

$$x \triangleq \alpha_1 u_* + \sum_{i=1}^{d_1 - 1} \beta_1^{(i)} \widetilde{u}_i,$$

$$y \triangleq \alpha_2 v_* + \sum_{j=1}^{d_2 - 1} \beta_2^{(j)} \widetilde{v}_j.$$

Substitute x and y in (7) and (8) by their expansion, we then have

$$
\begin{aligned}
(\alpha_1^2 + \sum_{i=1}^{d_1-1} (\beta_1^{(i)})^2)\alpha_2 &= \alpha_1, \\
(\alpha_2^2 + \sum_{i=1}^{d_2-1} (\beta_2^{(i)})^2)\alpha_1 &= \alpha_2, \\
(\alpha_1^2 + \sum_{i=1}^{d_1-1} (\beta_1^{(i)})^2)\beta_2^{(j)} &= 0, \forall j = 1, ..., d_2 - 1, \\
(\alpha_2^2 + \sum_{i=1}^{d_2-1} (\beta_2^{(i)})^2)\beta_1^{(j)} &= 0, \forall j = 1, ..., d_1 - 1.
\end{aligned}
$$

The above equalities yield the following two types of stationary points.

- $(\alpha_1^2 + \sum_{i=1}^{d_1-1}(\beta_1^{(i)})^2)(\alpha_2^2 + \sum_{i=1}^{d_2-1}(\beta_2^{(i)})^2) \neq 0, \beta_1 = 0, \beta_2 = 0$. This leads to $\alpha_1\alpha_2 = 1$. Thus, $xy^\top = \alpha_1\alpha_2 u_* v_*^\top = M$, and $\mathcal{F}(\alpha_1 u_*, \alpha_2 v_*) = 0$. Then we have global optima $\left(\alpha u_*, \frac{1}{\alpha} v_*\right)$ for $\alpha \neq 0$.

- Either $(\alpha_1^2 + \sum_{i=1}^{d_1-1}(\beta_1^{(i)})^2) = 0$ and $\alpha_2 = 0$, or $(\alpha_2^2 + \sum_{i=1}^{d_2-1}(\beta_2^{(i)})^2) = 0$ and $\alpha_1 = 0$. We next show that stationary points satisfy these conditions are strict saddle points. We only consider the first case, and the second case can be proved following similar lines. We first calculate the Hessian matrix as follows.

  $$
  \nabla^2 \mathcal{F}(x, y) = \begin{pmatrix} \|y\|_2^2 I_{d_1} & 2xy^\top - M \\ 2yx^\top - M^\top & \|x\|_2^2 I_{d_2} \end{pmatrix}.
  $$

  At $x = 0$, $y = \sum_{j=1}^{d_2-1} \beta_2^{(j)} \widetilde{v}_j$,

  $$
  \nabla^2 \mathcal{F}(x, y) = \begin{pmatrix} \sum_{i=1}^{d_2-1} (\beta_2^{(i)})^2 I_{d_1} & -M \\ -M^\top & 0 \end{pmatrix}.
  $$

  For any $a \in \mathbb{R}^{d_1}, b \in \mathbb{R}^{d_2}$,

  $$
  \left(a^\top, b^\top\right) \nabla^2 \mathcal{F}(x, y) \begin{pmatrix} a \\ b \end{pmatrix} = \sum_{i=1}^{d_2-1} (\beta_2^{(i)})^2 \|a\|_2^2 - 2a^\top M b
  $$

  For $(a, b) = (\widetilde{u}_i, \widetilde{v}_j)$, this quantity is positive. For $(a, b) = (u_*, \sum_{i=1}^{d_2-1}(\beta_2^{(i)})^2 v_*)$, this quantity is negative. Thus, $x = 0$, $y = \sum_{j=1}^{d_2-1} \beta_2^{(j)} \widetilde{v}_j$satisfies strict saddle property. We conclude that for any $x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}$ such that $x^\top u_* = y^\top v_* = 0$, we have strict saddle points $(x, 0)$ and $(0, y)$.

□

2.2 Proof of Lemma 2

At $x = \alpha u_*, y = \frac{1}{\alpha} v_*$,

$$
\nabla^2 \mathcal{F}(\alpha u_*, \frac{1}{\alpha} v_*) = \begin{pmatrix} \frac{1}{\alpha^2} I_{d_1} & M \\ M^\top & \alpha^2 I_{d_2} \end{pmatrix}.
$$

One can verify that $\nabla^2 \mathcal{F}\left(\alpha u_*, \frac{1}{\alpha} v_*\right)$ has eigenvalues $\alpha^2 + \frac{1}{\alpha^2}, \alpha^2, \frac{1}{\alpha^2}$. The largest eigenvalue is $\lambda_1 = \alpha^2 + \frac{1}{\alpha^2}$ and the smallest eigenvalue is $\lambda_{d_1+d_2} = \min\{\alpha^2, \frac{1}{\alpha^2}\}$. Thus, the condition number can be easily calculated as follows.

$$
\kappa\left(\nabla^2 \mathcal{F}\left(\alpha u_*, \frac{1}{\alpha} v_*\right)\right) = \max\{\alpha^4, \frac{1}{\alpha^4}\} + 1.
$$

2.3 Proof of Lemma 3

Proof. Recall that $M = u_* v_*^\top$ and

$$x \triangleq \alpha_1 u_* + \sum_{i=1}^{d_1-1} \beta_1^{(i)} \widetilde{u}_i,$$
$$y \triangleq \alpha_2 v_* + \sum_{j=1}^{d_2-1} \beta_2^{(j)} \widetilde{v}_j.$$

By setting the gradient of $\widetilde{\mathcal{F}}$ to zero we get

$$(\|x\|_2^2 + d_1\sigma_1^2) y = M^\top x,$$
$$(\|y\|_2^2 + d_2\sigma_2^2) x = My.$$

From the equations above, we can verify that $(x, y) = (0, 0)$ is a stationary point. Furthermore, left multiplying the equations above by $\widetilde{v}_j^\top$ and $\widetilde{u}_i^\top$ respectively, we will get that $\beta_1^{(i)}$'s and $\beta_2^{(j)}$'s are all zeros. Similarly, by left multiplying the equations above by $v_*^\top$ and $u_*^\top$ respectively, we will get

$$(\alpha_1^2 + d_1\sigma_1^2)\alpha_2^2 = \alpha_2\alpha_1,$$
$$(\alpha_2^2 + d_2\sigma_2^2)\alpha_1^2 = \alpha_1\alpha_2.$$

Then, with some algebraic manipulations, we get

$$\alpha_1^2 = \sqrt{\frac{d_1\sigma_1^2}{d_2\sigma_2^2}} - d_1\sigma_1^2,$$
$$\alpha_2^2 = \sqrt{\frac{d_2\sigma_2^2}{d_1\sigma_1^2}} - d_2\sigma_2^2.$$

Specifically, when $d_1\sigma_1^2 = \gamma^2 d_2\sigma_2^2 \le \gamma$, we have

$$\alpha_1 = \gamma\alpha_2 = \pm\sqrt{\gamma - \gamma^2 d_2\sigma_2^2}.$$

Next, we are going to show that $(0, 0)$ is a strict saddle point and $(x_*, y_*) \triangleq \pm(\alpha_1 u_*, \alpha_2 v_*)$ are global optima. We first calculate the Hessian matrix as follows.

$$\nabla^2 \widetilde{\mathcal{F}}(x, y) = \begin{pmatrix} \left(\|y\|_2^2 + d_2\sigma_2^2\right) I_{d_1} & 2xy^\top - M \\ 2yx^\top - M^\top & \left(\|x\|_2^2 + d_1\sigma_1^2\right) I_{d_2} \end{pmatrix}.$$

Since the injected noise is small, we have $\alpha_1\alpha_2 = 1 - \sqrt{d_1\sigma_1^2 d_2\sigma_2^2} > 0$. For any $a \in \mathbb{R}^{d_1}$ and $b \in \mathbb{R}^{d_2}$, we have

$$\begin{aligned}
&\left(a^\top, b^\top\right) \nabla^2 \widetilde{\mathcal{F}}(x_*, y_*) \begin{pmatrix} a \\ b \end{pmatrix} \\
=&(\alpha_2^2 + d_2\sigma_2^2)\|a\|_2^2 + (\alpha_1^2 + d_1\sigma_1^2)\|b\|_2^2 + 2(2\alpha_1\alpha_2 - 1)(a^\top u_*)(b^\top v_*) \\
\ge&(\alpha_2^2 + d_2\sigma_2^2)\|a\|_2^2 + (\alpha_1^2 + d_1\sigma_1^2)\|b\|_2^2 - 2|2(1 - \sqrt{d_1\sigma_1^2 d_2\sigma_2^2}) - 1|\ \|a\|_2\ \|b\|_2 \\
>&\sqrt{\frac{d_2\sigma_2^2}{d_1\sigma_1^2}}\|a\|_2^2 + \sqrt{\frac{d_1\sigma_1^2}{d_2\sigma_2^2}}\|b\|_2^2 - 2\|a\|_2\ \|b\|_2,
\end{aligned}$$

where the last inequality comes from the fact that $d_1\sigma_1^2$ and $d_2\sigma_2^2$ should be small enough such that

$$0 < 1 - \sqrt{d_1\sigma_1^2 d_2\sigma_2^2} < 1.$$

Note that

$$\sqrt{\frac{d_2\sigma_2^2}{d_1\sigma_1^2}}\|a\|_2^2+\sqrt{\frac{d_1\sigma_1^2}{d_2\sigma_2^2}}\|b\|_2^2-2\|a\|_2\ \|b\|_2=\left((\frac{d_2\sigma_2^2}{d_1\sigma_1^2})^{\frac{1}{4}}\|a\|_2-(\frac{d_1\sigma_1^2}{d_2\sigma_2^2})^{\frac{1}{4}}\|b\|_2\right)^2.$$

Thus, as long as $d_1\sigma_1^2\ d_2\sigma_2^2<1$, we have $\nabla^2\mathcal{F}_{reg}(x_*,y_*)$ is positive definite (PD). Thus (x_*,y_*) is global minimum and so is $(-x_*,-y_*)$.

Similarly, for any $a\in\mathbb{R}^{d_1}$ and $b\in\mathbb{R}^{d_2}$, we have

$$\left(a^\top,b^\top\right)\nabla^2\mathcal{F}_{reg}(0,0)\begin{pmatrix}a\\b\end{pmatrix}=d_2\sigma_2^2\|a\|_2^2+d_1\sigma_1^2\|b\|_2^2-2(a^\top u_*)(b^\top v_*).$$

It is easy to show that for $(a,b)=(u_*,v_*)$, the quantity is negative when noise is small enough. But for $(a,b)=(\widetilde{u}_i,\widetilde{v}_j)$, this quantity is positive. Thus, (0,0) is a strict saddle point. Here, we prove Lemma 3. □

3 Proof of Theorem 1

3.1 Boundedness of Trajectory

We first show that the solution trajectory of Perturbed GD is bounded with high probability, which is a sufficient condition for our following convergence analysis.

Lemma 1 (Boundedness of Trajectories)**.** Given $x_0\in\mathbb{R}^{d_1}$, $y_0\in\mathbb{R}^{d_2}$, we choose $\sigma_1,\sigma_2>0$ such that $\sigma^2=\mathbb{E}\left[\|\xi_1\|_2^2\right]=\mathbb{E}\left[\|\xi_2\|_2^2\right]$ and $\|x_0\|_2^2+\|y_0\|_2^2\le 1/\sigma^2$. For any $\delta\in(0,1)$, we take

$$\eta\le\eta_1=C_1\sigma^6(\log((d_1+d_2)/\delta)\log(1/\delta))^{-1},$$

for some positive constant C_1. Then with probability at least $1-\delta$, we have $\|x_t\|_2^2+\|y_t\|_2^2\le 2/\sigma^2$ for any $t\le T_1=O(1/\eta^2)$.

Proof. We first define the event where the injected noise for both x and y is bounded for the first t iterations.

$$\mathcal{A}_t=\left\{|\xi_{1,\tau}^{(i)}|,|\xi_{2,\tau}^{(j)}|\le\sigma\left(\sqrt{2\log((d_1+d_2)\eta^{-2})}+\sqrt{\log(1/\delta)}\right),\forall\tau\le t,i=1,...,d_1,j=1,..,d_2\right\}.\tag{9}$$

By the concentration result of the maximum of Gaussian distribution, we have $\mathbb{P}(\mathcal{A}_{1/\eta^2})\ge 1-\delta$. Moreover, we use $\mathcal{H}_t$ to denote the event where the first t iterates $\{(x_\tau,y_\tau)\}_{\tau\le t}$ is bounded, i.e., $\mathcal{H}_t=\left\{\|x_\tau\|_2^2+\|y_\tau\|_2^2\le\frac{2}{\sigma_2},\forall\tau\le t\right\}$. Let $\mathcal{F}_t=\sigma\{(x_\tau,y_\tau),\tau\le t\}$ denote the $\sigma-$algebra generated by that past t iterations.

Under the event $\mathcal{H}_t$ and $\mathcal{A}_t$, we have the following inequality on the conditional expectation of $\|x_{t+1}\|_2^2+\|y_{t+1}\|_2^2$.

$$\begin{aligned}
&\mathbb{E}\left[(\|x_{t+1}\|_2^2+\|y_{t+1}\|_2^2)\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\middle|\mathcal{F}_t\right]\\
=&\left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right)-4\eta\left(\|x_t\|_2^2\|y_t\|_2^2-x_t^\top My_t\right)\right\}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&+\eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\|\nabla_x\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t})\|_2^2+\|\nabla_y\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t})\|_2^2\right]\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
\le&\left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right)-4\eta\left(\|x_t\|_2^2\|y_t\|_2^2-\|x_t\|_2\|y_t\|_2\right)\right\}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&+\eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\left|\|y_t+\xi_{2,t}\|_2^2(x_t+\xi_{1,t})-M(y_t+\xi_{2,t})\right|_2^2\right]\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&+\eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\left|\|x_t+\xi_{1,t}\|_2^2(y_t+\xi_{2,t})-M^\top(x_t+\xi_{1,t})\right|_2^2\right]\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
\le&\left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right)-4\eta\left(\|x_t\|_2^2\|y_t\|_2^2-\|x_t\|_2\|y_t\|_2\right)\right\}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&+2\eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\|y_t+\xi_{2,t}\|_2^4\|x_t+\xi_{1,t}\|_2^2+\|M(y_t+\xi_{2,t})\|_2^2\right]\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&+2\eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\|x_t+\xi_{1,t}\|_2^4\|y_t+\xi_{2,t}\|_2^2+\|M^\top(x_t+\xi_{1,t})\|_2^2\right]\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
\le&\left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right)+\eta+\eta^2C_2\right\}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t},
\end{aligned}$$

where the second inequality comes from the fact that $x^2-x\ge-\frac{1}{4}$ for all $x\in\mathbb{R}$ and $C_2=(\sigma^2+1/\sigma^2)(2/\sigma^4+6/d_1+6/d_2+6\sigma^4)$ in the last inequality. We take $\eta\le 1/C_2$, then we have

$$\begin{aligned}
\mathbb{E}\left[(\|x_{t+1}\|_2^2+\|y_{t+1}\|_2^2-1/\sigma^2)\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\middle|\mathcal{F}_t\right]&\le(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2-1/\sigma^2\right)\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\\
&\le(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2-1/\sigma^2\right)\mathbb{1}_{\mathcal{H}_{t-1}\cap\mathcal{A}_{t-1}}.
\end{aligned}\tag{10}$$

If we denote $G_t = (1-2\eta\sigma^2)^{-t}(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2 - 1/\sigma^2)$, $G_t \mathbb{1}_{\mathcal{H}_{t-1}\cap\mathcal{A}_{t-1}}$ is then a super-martingale according to (10). We will apply Azuma's Inequality to prove the bound and before that we have to bound the difference between $G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}$ and $\mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}|\mathcal{F}_t]$.

$$
\begin{aligned}
d_{t+1} &= \left|G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}|\mathcal{F}_t]\right| \\
&= (1-2\eta\sigma^2)^{-t-1}\Big|2\eta x_t^\top[(x_t+\xi_{1,t})(y_t+\xi_{2,t})^\top - M](y_t+\xi_{2,t})) \\
&\qquad - 2\eta\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[x_t^\top[(x_t+\xi_{1,t})(y_t+\xi_{2,t})^\top - M](y_t+\xi_{2,t}))\right] \\
&\qquad + 2\eta y_t^\top[(x_t+\xi_{1,t})(y_t+\xi_{2,t})^\top - M]^\top(x_t+\xi_{1,t})) \\
&\qquad - 2\eta\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[y_t^\top[(x_t+\xi_{1,t})(y_t+\xi_{2,t})^\top - M]^\top(x_t+\xi_{1,t}))\right] \\
&\qquad + \eta^2\left|\left|\|y_t+\xi_{2,t}\|_2^2(x_t+\xi_{1,t}) - M(y_t+\xi_{2,t})\right|\right|_2^2 \\
&\qquad - \eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\left|\left|\|y_t+\xi_{2,t}\|_2^2(x_t+\xi_{1,t}) - M(y_t+\xi_{2,t})\right|\right|_2^2\right] \\
&\qquad + \eta^2\left|\left|\|x_t+\xi_{1,t}\|_2^2(y_t+\xi_{2,t}) - M^\top(x_t+\xi_{1,t})\right|\right|_2^2 \\
&\qquad - \eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\left|\left|\|x_t+\xi_{1,t}\|_2^2(y_t+\xi_{2,t}) - M^\top(x_t+\xi_{1,t})\right|\right|_2^2\right]\Big|\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t} \\
&\le C_1'(1-2\eta\sigma^2)^{-t-1}\eta\sigma^{-2}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right),
\end{aligned}
$$

where C_1' is some positive constant. Denote $r_t = \sqrt{\sum_{i=1}^t d_i^2}$. By Azuma's inequality, we have

$$
\mathbb{P}\left(G_t\mathbb{1}_{\mathcal{H}_{t-1}\cap\mathcal{A}_{t-1}} - G_0 \ge O(1)r_t\left(\log\frac{1}{\eta^2\delta}\right)^{\frac{1}{2}}\right) \le O(\eta^2\delta).
$$

Then when with probability at least $1-O(\eta^2\delta)$, we have

$$
\begin{aligned}
(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2)\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t} &\le 1/\sigma^2 + (1-\eta\sigma^2)^t(\|x_0\|_2^2 + \|y_0\|_2^2 - 1/\sigma^2) \\
&\quad + O(1)(1-\eta\sigma^2)^t r_t(\log\frac{1}{\eta^2\delta})^{\frac{1}{2}} \\
&\le 1/\sigma^2 + 0 + O\left(\sqrt{\eta}\sigma^{-2}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\right)\left(\log\frac{1}{\eta^2\delta}\right)^{\frac{1}{2}} \\
&\le 2/\sigma^2,
\end{aligned}
$$

when $\eta = O\left(\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1}\right)$. In order to satisfy $\eta \le 1/C_2$ at the same time, we take $\eta \le \eta_1 = O\left(\sigma^6\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1}\right)$ to make sure that all inequalities above hold.

The above inequality shows that if $\mathcal{H}_{t-1}\cap\mathcal{A}_{t-1}$ holds, then $\mathcal{H}_t\cap\mathcal{A}_t$ holds with probability at least $1-O(\eta^2\delta)$. Hence with probability at least $1-\delta$, we have $(\|x_t\|_2^2 + \|y_t\|_2^2)\mathbb{1}_{\mathcal{A}_{t-1}} \le 2/\sigma^2$ for all $t \le T_1 = O(1/\eta^2)$. Recall that

$$
\mathbb{P}(\mathcal{A}_{1/\eta^2}) \ge 1-\delta.
$$

Thus, we have with probability at least $1-2\delta$, $\|x_t\|_2^2 + \|y_t\|_2^2 \le 2/\sigma^2$ for all $t \le T_1 = O(1/\eta^2)$. By properly rescaling δ, we prove Lemma 1. □

3.2 Proof of Lemma 4

Proof. We only prove the convergence of $\|\beta_{1,t}\|_2^2$ here. The proof of the convergence of $\|\beta_{2,t}\|_2^2$ follows similar lines. For notational simplicity, we denote $\xi_{1,t}^{(-1)} = (\xi_{1,t}^{(2)}, ..., \xi_{1,t}^{(d_1)})^\top$, $\forall t \ge 0$. We first bound the conditional expectation

of $\|\beta_{1,t+1}\|_2^2$ given $\mathcal{F}_t$:

$$\begin{aligned}\mathbb{E}\left[\|\beta_{1,t+1}\|_2^2\big|\mathcal{F}_t\right] &= \|\beta_{1,t}\|_2^2 + \eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\|y_t+\xi_{2,t}\|_2^4\|\beta_{1,t}+\xi_{1,t}^{(-1)}\|_2^2\right] - 2\eta\left(\|y_t\|_2^2+\sigma^2\right)\|\beta_{1,t}\|_2^2 \\ &\leq (1-2\eta\sigma^2)\|\beta_{1,t}\|_2^2 + \eta^2 C_2, \end{aligned} \tag{11}$$

where $C_2 = (\sigma^2+1/\sigma^2)(2/\sigma^4+6/d_1+6/d_2+6\sigma^4)$. The bound on the second moment comes from the boundedness of $\|y_t\|_2^2$, which has been shown in Lemma 1. Define $G_t = (1-2\eta\sigma^2)^{-t}\left(\|\beta_{1,t}\|_2^2 - \frac{\eta C_2}{2\sigma^2}\right)$, and $\mathcal{E}_t = \left\{\forall\tau\leq t, \|\beta_{1,\tau}\|_2^2 \geq \frac{\eta C_2}{\sigma^2}\right\}$. By (11), we have

$$\mathbb{E}\left[G_{t+1}\mathbb{1}_{\mathcal{E}_t}\big|\mathcal{F}_t\right] \leq G_t\mathbb{1}_{\mathcal{E}_t} \leq G_t\mathbb{1}_{\mathcal{E}_{t-1}}.$$

Hence, by Markov inequality we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_t) = \mathbb{P}\left(\|\beta_{1,t}\|_2^2\mathbb{1}_{\mathcal{E}_{t-1}} \geq \frac{\eta C_2}{\sigma^2}\right) &\leq \frac{\mathbb{E}\left[\|\beta_{1,t}\|_2^2\mathbb{1}_{\mathcal{E}_{t-1}}\right]}{\frac{\eta C_2}{\sigma^2}} \\ &\leq \frac{(1-2\eta\sigma^2)^t(\|\beta_{1,0}\|_2^2 - \frac{\eta C_2}{2\sigma^2}) + \frac{\eta C_2}{2\sigma^2}}{\frac{\eta C_2}{\sigma^2}} \\ &\leq (1-2\eta\sigma^2)^t\frac{2}{C_2\eta} + \frac{1}{2} \leq \frac{3}{4},\end{aligned}$$

when $t \geq \frac{1}{2\eta\sigma^2}\log\frac{8}{C_2\eta}$. We take $t = \frac{1}{\eta\sigma^2}\log\frac{8}{C_2\eta}$ to make sure the inequality above holds. Thus with probability at least $\frac{1}{4}$, there exists a $\tau \leq \frac{1}{\eta\sigma^2}\log\frac{8}{C_2\eta}$, such that $\|\beta_{1,\tau}\|_2^2 \leq \frac{\eta C_2}{\sigma^2}$. Thus, with probability at least $1-\delta$, we can find a τ such that $\|\beta_{1,\tau}\|_2^2 \leq \frac{\eta C_2}{\sigma^2}$, where

$$\tau \leq \tau_1 = \frac{1}{\log(4/3)\eta\sigma^2}\log\frac{8}{C_2\eta}\log\frac{1}{\delta} = \frac{1}{\log(4/3)\eta\sigma^2}\left(\log\frac{8}{C_2}+\log\frac{1}{\eta}\right)\log\frac{1}{\delta} = O\left(\frac{1}{\eta\sigma^2}\log\frac{1}{\eta}\log\frac{1}{\delta}\right).$$

We next show that with probability at least $1-\delta$, for $\forall t \geq \tau_1$, we have $\|\beta_{1,t}\|_2^2 \leq 2\frac{\eta C_2}{\sigma^2}$. This can be done following the similar lines to the proof of Lemma 1. We first restart the counter of the time and assume $\|\beta_{1,0}\|_2^2 \leq \frac{\eta C_2}{\sigma^2}$. Denote $\mathcal{H}_t = \left\{\forall\tau\leq t, \|\beta_{1,\tau}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2}\right\}$. Then by (11), we have

$$\mathbb{E}\left[G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}\big|\mathcal{F}_t\right] \leq G_t\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t} \leq G_t\mathbb{1}_{\mathcal{H}_{t-1}\cap\mathcal{A}_{t-1}}.$$

The difference of $G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}$ and $\mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}|\mathcal{F}_t]$ can be easily bounded as follows

$$D_{t+1} = \left|G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{H}_t\cap\mathcal{A}_t}|\mathcal{F}_t]\right| = C_2'\eta^2\sigma^{-3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right),$$

where C_2' is some constant. Then by applying Azuma's inequality and following the similar lines to the proof of Lemma 1, we show that when

$$\eta \leq C_3'\sigma^8\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1},$$

with probability at least $1-\delta$, we have

$$\|\beta_{1,t}\|_2^2 \leq 2\frac{\eta C_2}{\sigma^2}.$$

for any $\tau_1 \leq t \leq T_1 = O(1/\eta^2)$. □

3.3 Proof of Lemma 5

We prove this lemma in three steps.

• **Step 1:** The following lemma shows that after polynomial time, with high probability, the algorithm can move out of the $O(\eta)$ neighborhood of the saddle point.

Lemma 2 (Escaping from the Unique Saddle Point)**.** Suppose $\|\beta_{1,t}\|_2^2 \le 2\frac{\eta C_2}{\sigma^2}$ and $\|\beta_{2,t}\|_2^2 \le 2\frac{\eta C_2}{\sigma^2}$ hold for all $t > 0$. For $\forall \delta \in (0,1)$, we take

$$\eta = O\left(\sigma^{12}\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1}\right).$$

Then with probability at least $1-\delta$, there exists a $\tau \le \tau_2$, such that

$$x_\tau^\top M y_\tau \ge 9\frac{\eta C_2}{\sigma^4},$$

where $\tau_2 = \frac{5}{\eta\sigma^2}\log\frac{4\sigma^3}{\eta C_2}\log\frac{1}{\delta}$.

Proof. Let's define the event $\mathcal{H}_t = \{x_\tau^\top M y_\tau \le 9\frac{\eta C_2}{\sigma^4},\ \forall \tau \le t\}$. Following the proof of Lemma 1, we can refine the bound on the conditional expectation of $\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2$ given $\mathcal{H}_t$. Specifically, we have

$$\begin{aligned}
\mathbb{E}\left[(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2)\mathbb{1}_{\mathcal{H}_t}\middle|\mathcal{F}_t\right] &= \left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right) - 4\eta\left(\|x_t\|_2^2\|y_t\|_2^2 - x_t^\top M y_t\right)\right\}\mathbb{1}_{\mathcal{H}_t} \\
&\quad + \eta^2\mathbb{E}_{\xi_{1,t},\xi_{2,t}}\left[\|\nabla_x\mathcal{F}(x_t+\xi_{1,t}, y_t+\xi_{2,t})\|_2^2 + \|\nabla_y\mathcal{F}(x_t+\xi_{1,t}, y_t+\xi_{2,t})\|_2^2\right]\mathbb{1}_{\mathcal{H}_t} \\
&\le \left\{(1-2\eta\sigma^2)\left(\|x_t\|_2^2+\|y_t\|_2^2\right) + 34C_2\eta^2\sigma^{-4})\right\}\mathbb{1}_{\mathcal{H}_t}.
\end{aligned}$$

Denote $G_t = (1-2\eta\sigma^2)^{-t}\left(\|x_t\|_2^2 + \|y_t\|_2^2 - 17C_2\eta\sigma^{-6}\right)$. Then we have

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{H}_t}] \le G_t\mathbb{1}_{\mathcal{H}_t} \le G_t\mathbb{1}_{\mathcal{H}_{t-1}}.$$

Thus, by Markov inequality, we have

$$\begin{aligned}
\mathbb{P}((\|x_t\|_2^2+\|y_t\|_2^2)\mathbb{1}_{\mathcal{H}_{t-1}} \ge 34C_2\eta\sigma^{-6}) &\le \frac{(1-2\eta\sigma^2)^t(\|x_0\|_2^2+\|y_0\|_2^2 - 17C_2\eta\sigma^{-6}) + 17C_2\eta\sigma^{-6}}{34C_2\eta\sigma^{-6}} \\
&\le \frac{3}{4},
\end{aligned}$$

when $t \ge \frac{1}{2\eta\sigma^2}\log\frac{4\sigma^3}{\eta C_2}$. Thus with probability at least $1-\delta$, there exists a $\tau \le \tau_2 = \frac{5}{2\eta\sigma^2}\log\frac{4\sigma^3}{\eta C_2}\log\frac{1}{\delta} = O(\frac{1}{\eta\sigma^2}\log\frac{1}{\eta}\log\frac{1}{\delta})$, such that

$$(\|x_\tau\|_2^2 + \|y_\tau\|_2^2)\mathbb{1}_{\mathcal{H}_{\tau-1}} \le 34C_2\eta\sigma^{-6}.$$

Following the exactly same proof of Lemma 1, we can show that with probability at least $1-\delta$, for all $\tau_2 \le t \le T_1 = O(\frac{1}{\eta^2})$, we have

$$\begin{aligned}
(\|x_t\|_2^2+\|y_t\|_2^2)\mathbb{1}_{\mathcal{H}_{t-1}} &\le (17+34)C_2\eta\sigma^{-6} + O\left(\sqrt{\eta}\sigma^{-2}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\right)\left(\log\frac{1}{\eta^2\delta}\right)^{\frac{1}{2}} \\
&= C_4'\eta\sigma^{-12} + C_5'\sqrt{\eta}\sigma^{-2}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\left(\log\frac{1}{\eta^2\delta}\right)^{\frac{1}{2}} \\
&\le C_6'\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1} + C_7'\sigma^4,
\end{aligned}$$

where C_4', C_5', C_6' and C_7' are some positive constants, and when

$$\eta = O\left(\sigma^{12}\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1}\right).$$

We next show that for large enough t, with high probability, $\mathcal{H}_{t-1}$ does not hold. We prove by contradiction: if $\mathcal{H}_{t-1}$ holds for all t, the solution trajectory stays in a small neighborhood around 0. If so, we can show that with constant probability, $|\alpha_{1,t}+\alpha_{2,t}|$ will explode to infinity, which is in contradiction with the boundedness. Here follows the detailed proof.

Assuming $\mathcal{H}_{t-1}$ holds for all $t \le \widetilde{\tau}_2 = O(\frac{1}{\eta}\log\frac{1}{\eta\sigma}\log\frac{1}{\delta})$, by the analysis above we have $\|x_t\|_2^2 + \|y_t\|_2^2 \le C_6'\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1} + C_7'\sigma^4$ holds for $\forall t \le \widetilde{\tau}_2$.

Note that with at least some constant probability,

$$\begin{aligned}
&\left|\alpha_{1,t+1}+\alpha_{2,t+1}\right|-\left|\alpha_{1,t}+\alpha_{2,t}\right| \\
=&\left|\alpha_{1,t}+\alpha_{2,t}-\eta\left(<\nabla_x\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t}),u_*>+<\nabla_y\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t}),v_*>\right)\right|-\left|\alpha_{1,t}+\alpha_{2,t}\right| \\
\geq& C_8{}'\eta(\sigma_1+\sigma_2),
\end{aligned}$$

where $C_8{}'$ is some positive constant. This means we can find a $\tau = O(\log\frac{1}{\delta}) \leq \widetilde{\tau}_2$, such that $\left|\alpha_{1,\tau}+\alpha_{2,\tau}\right| > C_8{}'\eta(\sigma_1+\sigma_2)$, with probability at least $1-\delta$. We will use this point as our initialization for the following proof. We next give a lower bound on the conditional expectation of $|\alpha_{1,t+1}+\alpha_{2,t+1}|$.

$$\begin{aligned}
&\mathbb{E}\left[\left|\alpha_{1,t+1}+\alpha_{2,t+1}\right|\middle|\mathcal{F}_t\right] \\
\geq&\left|\mathbb{E}\left[\left(\alpha_{1,t+1}+\alpha_{2,t+1}\right)\middle|\mathcal{F}_t\right]\right| \\
=&\left|\left(1+\eta(1-\sigma^2-\alpha_{1,t}\alpha_{2,t})\right)\left(\alpha_{1,t}+\alpha_{2,t}\right)-\eta\left(\|\beta_{1,t}\|_2^2\alpha_{2,t}+\|\beta_{2,t}\|_2^2\alpha_{1,t}\right)\right| \\
\geq&\left(1+\eta(1-\sigma^2-\alpha_{1,t}\alpha_{2,t})\right)\left|\alpha_{1,t}+\alpha_{2,t}\right|-2C_9{}'\frac{\eta^{2.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right) \\
\geq&\left(1+\frac{1}{2}\eta\right)\left|\alpha_{1,t}+\alpha_{2,t}\right|-2C_9{}'\frac{\eta^{2.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right),
\end{aligned}$$

where $C_9{}'$ is some positive constant. This is equivalent to the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\left|\alpha_{1,t+1}+\alpha_{2,t+1}\right|-4C_9{}'\frac{\eta^{1.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\middle|\mathcal{F}_t\right] \\
\geq&\left(1+\frac{1}{2}\eta\right)\left(\left|\alpha_{1,t}+\alpha_{2,t}\right|-4C_9{}'\frac{\eta^{1.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\right) \\
\geq&\left(1+\frac{1}{2}\eta\right)^{t+1}\left(\left|\alpha_{1,0}+\alpha_{2,0}\right|-4C_9{}'\frac{\eta^{1.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\right) \\
\geq&\left(1+\frac{1}{2}\eta\right)^{t+1}\left(C_8{}'\eta(d_1+d_2)-4C_9{}'\frac{\eta^{1.2}C_2}{\sigma^3}\left(\left(\log\frac{\sigma_1+\sigma_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)\right) \\
\geq& C_{10}{}'\eta(\sigma_1+\sigma_2)\exp t\eta/2 \\
\geq& C_{10}{}'\eta(\sigma/\sqrt{d_1}+\sigma/\sqrt{d_2})\frac{1}{\eta\sigma} \\
=& C_{10}{}'(1/\sqrt{d_1}+1/\sqrt{d_2}),
\end{aligned}$$

where $C_{10}{}'$ is some positive constant and last inequality holds when $t \geq \frac{2}{\eta}\log\frac{1}{\eta\sigma}$. Note that here we still take $\eta = O\left(\sigma^{12}\left(\log\frac{d_1+d_2}{\delta}\log\frac{1}{\delta}\right)^{-1}\right)$, which makes sure that $\frac{\eta^{1.2}C_2}{\sigma^3}\left(\left(\log\frac{d_1+d_2}{\eta}\right)^{\frac{1}{2}}+\left(\log\frac{1}{\delta}\right)^{\frac{1}{2}}\right)$ is small enough to make the fourth inequality hold.

For fixed d_1 and d_2, if we let δ and σ go to zero (which guarantees that η goes to zero), we will have the conditional expectation of $|\alpha_{1,t+1}+\alpha_{2,t+1}|$ stays in a neighborhood around a positive constant. However, by our assumption, $\|x_{t+1}\|_2^2+\|y_{t+1}\|_2^2$ stays in a very small neighborhood around zero, which makes $\alpha_{1,t+1}^2+\alpha_{2,t+1}^2 \leq \|x_{t+1}\|_2^2+\|y_{t+1}\|_2^2$ also very small. This implies that $|\alpha_{1,t+1}+\alpha_{2,t+1}|$ can be arbitrarily small as long as we make δ and σ small enough, which is the desired contradiction.

Thus, we know that, with probability at least $1-\delta$, there exists a t satisfying $\tau_2 \leq t \leq \tau_2+\widetilde{\tau}_2 \leq 2\tau_2$, such that $\mathcal{H}_t$ does not hold. Re-scale τ_2 and we prove the result. □

• **Step 2:** The following lemma shows that after step 1, with high probability, the algorithm will continue move away from the saddle point and escape from the saddle point at some time.

Lemma 3. Suppose $x_0^\top M y_0 \geq 9\frac{\eta C_2}{\sigma^4}$. We take η as in Lemma 5. Then there almost surely exists a $\tau \leq \tau_2' = \frac{2}{\eta}\log\frac{2\sigma^3}{\eta C_2}$, such that

$$x_\tau^\top M y_\tau \geq \frac{1}{2} + \sigma^2.$$

Proof. Suppose $x_t^\top M y_t \leq \frac{1}{2} + \sigma^2$ holds for all $t \leq \tau_2'$. We next give a lower bound on the conditional expectation of $|\alpha_{1,t+1} + \alpha_{2,t+1}|$.

$$\begin{aligned}
\mathbb{E}\left[|\alpha_{1,t+1} + \alpha_{2,t+1}||\mathcal{F}_t\right] &\geq \left|\mathbb{E}\left[(\alpha_{1,t+1} + \alpha_{2,t+1})|\mathcal{F}_t\right]\right| \\
&= \left|\left(1 + \eta(1 - \sigma^2 - \alpha_{1,t}\alpha_{2,t})\right)(\alpha_{1,t} + \alpha_{2,t}) - \eta\left(\|\beta_{1,t}\|_2^2\alpha_{2,t} + \|\beta_{2,t}\|_2^2\alpha_{1,t}\right)\right| \\
&\geq \left(1 + \frac{1}{2}\eta\right)\left|\alpha_{1,t} + \alpha_{2,t}\right| - 4\frac{\eta^2 C_2}{\sigma^4}.
\end{aligned}$$

This implies that

$$\mathbb{E}\left[|\alpha_{1,t+1} + \alpha_{2,t+1}| - 8\frac{\eta C_2}{\sigma^4}|\mathcal{F}_t\right] \geq \left(1 + \frac{1}{2}\eta\right)\left(\left|\alpha_{1,t} + \alpha_{2,t}\right| - 8\frac{\eta C_2}{\sigma^4}\right).$$

The above inequality further implies a lower bound on the conditional expectation of $|\alpha_{1,t+1} + \alpha_{2,t+1}| - 8\frac{\eta C_2}{\sigma^4}$.

$$\begin{aligned}
\mathbb{E}\left[|\alpha_{1,t+1} + \alpha_{2,t+1}| - 8\frac{\eta C_2}{\sigma^4}\right] &\geq \left(1 + \frac{1}{2}\eta\right)^t\left(\left|\alpha_{1,0} + \alpha_{2,0}\right| - 8\frac{\eta C_2}{\sigma^4}\right) \\
&\geq \left(1 + \frac{1}{2}\eta\right)^t\frac{\eta C_2}{\sigma^4} \geq \frac{2}{\sigma},
\end{aligned}$$

when $t = \tau_2' = \frac{2}{\eta}\log\frac{2\sigma^3}{\eta C_2}$. On the other hand,

$$\mathbb{E}\left[|\alpha_{1,t+1} + \alpha_{2,t+1}| - 8\frac{\eta C_2}{\sigma^4}\right] \leq \frac{\sqrt{2}}{\sigma},$$

which leads to a contradiction unless

$$\mathbb{P}(x_t^\top M y_t \leq \frac{1}{2} + \sigma^2,\ \forall t \leq \tau_2') = 0.$$

We prove the result. □

• **Step 3:** We then show that the algorithm will never iterate back towards the saddle point after escaping from it. Then Lemma 5 is proved.

Lemma 4. Suppose there exists a time step τ such that for some positive constant $c < \frac{1}{2}$

$$x_\tau^\top M y_\tau > 2c,$$

then $\forall\delta \in (0,1)$, with probability at least $1-\delta$, $\forall\tau \leq t \leq T_1 \triangleq O(\frac{1}{\eta^2})$,

$$(x_t, y_t) \in \left\{(x,y)|x^\top M y > c\right\}.$$

By taking $c = \frac{1}{4}$, we can prove Lemma 5 from Lemma 4.

Proof. We consider two cases:
(i) if $\alpha_{1,t}\alpha_{2,t} \geq 2(1-c)$, then $\alpha_{1,t+1}\alpha_{2,t+1} > 2c$ w.p.1.
(ii) if $\alpha_{1,t}\alpha_{2,t} < 2(1-c)$, then, $\forall\delta \in (0,1)$ and $\forall t \leq T_1 = O(\frac{1}{\eta^2})$, w.p. at least $1-\delta$,

$$(\alpha_{1,t}, \alpha_{2,t}) \in \{(\alpha_1, \alpha_2)|\left(\alpha_1\alpha_2 - 1\right)^2 < (1-c)^2\}.$$

Note that $(\alpha_{1,t}\alpha_{2,t}-1)^2<(1-c)^2$ implies that $\alpha_{1,t}\alpha_{2,t}>c$, thus we can prove Lemma 4. Further note that the injected noise is upper bounded with high probability, we assume in the following $T_1=O(\frac{1}{\eta^2})$ steps,

$$\|\xi_{k,t}\|_\infty\le\bar{\sigma}\triangleq\sigma\left(\sqrt{2\log((d_1+d_2)\eta^{-2})}+\sqrt{\log(1/\delta)}\right).$$

Case (i): By plugging (5) and (6) to (2), we get

$$\begin{aligned}\alpha_{1,t+1}\alpha_{2,t+1}&=\alpha_{1,t}\alpha_{2,t}-\eta\left(\alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2+\|\beta_{2,t}\|_2^2)-\alpha_{2,t}+g_x)+\alpha_{1,t}(\alpha_{2,t}(\alpha_{1,t}^2+\|\beta_{1,t}\|_2^2)-\alpha_{1,t}+g_y)\right)\\&\quad+\eta^2(\alpha_{1,t}(\alpha_{2,t}^2+\|\beta_{2,t}\|_2^2)-\alpha_{2,t}+g_x)(\alpha_{2,t}(\alpha_{1,t}^2+\|\beta_{1,t}\|_2^2)-\alpha_{1,t}+g_y)\\&=2c+(\alpha_{1,t}\alpha_{2,t}-2c)-\eta[\alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2+\|\beta_{2,t}\|_2^2)-\alpha_{2,t}+g_x)+\alpha_{1,t}(\alpha_{2,t}(\alpha_{1,t}^2+\|\beta_{1,t}\|_2^2)-\alpha_{1,t}+g_y)\\&\quad+\eta(\alpha_{1,t}(\alpha_{2,t}^2+\|\beta_{2,t}\|_2^2)-\alpha_{2,t}+g_x)(\alpha_{2,t}(\alpha_{1,t}^2+\|\beta_{1,t}\|_2^2)-\alpha_{1,t}+g_y)],\end{aligned}$$

First note that by Lemma 3 we can always find such $0<c<\frac{1}{2}$ in the condition. As $\alpha_{1,t}\alpha_{2,t}-2c>2(1-c)-2c=2(1-2c)>0$, we can prove (i) by choosing η small enough.

By Lemma 1, $\alpha_{1,t}^2+\alpha_{2,t}^2$ is bounded by $\frac{2}{\sigma^2}$ and we can know g_x and g_y are at most of order $O\left(\frac{1}{\sigma}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)\right)$ from the following upper bound:

$$\begin{aligned}|g_x|&\le\bar{\sigma}(\alpha_{2,t}^2+2\alpha_{2,t}\alpha_{1,t}+1+\frac{2\eta C_2}{\sigma^2})+\bar{\sigma}^2(2\alpha_{2,t}+\alpha_{1,t})+\bar{\sigma}^3,\\|g_y|&\le\bar{\sigma}(\alpha_{1,t}^2+2\alpha_{1,t}\alpha_{2,t}+1+\frac{2\eta C_2}{\sigma^2})+\bar{\sigma}^2(2\alpha_{1,t}+\alpha_{2,t})+\bar{\sigma}^3.\end{aligned}$$

It is easy to show that for properly selected

$$\eta\le\widetilde{\eta}_1=O\left(\left(\frac{1}{\sigma^4}+\frac{1}{\sigma^2}\left(\sqrt{\log(d_1+d_2)}+\sqrt{\log\frac{1}{\delta}}\right)\right)^{-1}\right),$$

the last two terms are greater than zero. Thus, w.p.1, $\alpha_{1,t+1}\alpha_{2,t+1}>2c$.
Case (ii): Without loss of generality, we place the time origin at t, and thus we have $2c<\alpha_1^0\alpha_2^0<2-2c$.

$$\mathbb{E}_\xi[A_t]=\mathbb{E}_\xi[\alpha_{2,t}<\nabla_x\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t}),u_*>+\alpha_{1,t}<\nabla_y\mathcal{F}(x_t+\xi_{1,t},y_t+\xi_{2,t}),v_*>].$$

By plugging (5) and (6) in the above equation, we have

$$\begin{aligned}\mathbb{E}_\xi[A_t]&=\alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2+\|\beta_{2,t}\|_2^2)-\alpha_{2,t}+\alpha_{1,t}\sigma^2)+\alpha_{1,t}(\alpha_{2,t}(\alpha_{1,t}^2+\|\beta_{1,t}\|_2^2)-\alpha_{1,t}+\alpha_{2,t}\sigma^2)\\&=(\alpha_{2,t}\alpha_{1,t}-1)(\alpha_{2,t}^2+\alpha_{1,t}^2)+\alpha_{2,t}\alpha_{1,t}(2\sigma^2+\|\beta_{1,t}\|_2^2+\|\beta_{2,t}\|_2^2).\end{aligned}$$

Again, by plugging (5) and (6) to (3) and taking expectation conditioning on $\mathcal{F}_t$, when $\alpha_{1,t}\alpha_{2,t}>c$, we will get

$$\begin{aligned}\mathbb{E}[(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2|\mathcal{F}_t]&=(\alpha_{1,t}\alpha_{2,t}-1)^2-2\eta\mathbb{E}_\xi[A_t](\alpha_{1,t}\alpha_{2,t}-1)\\&\quad+\eta^2\mathbb{E}_\xi[A_t^2]+\eta^4\mathbb{E}_\xi[B_t^2]-2\eta^3\mathbb{E}_\xi[A_tB_t]+2\eta^2\mathbb{E}_\xi[B_t](\alpha_{1,t}\alpha_{2,t}-1)\\&=\left(1-2\eta(\alpha_{2,t}^2+\alpha_{1,t}^2)\right)(\alpha_{1,t}\alpha_{2,t}-1)^2\\&\quad-2\eta\alpha_{2,t}\alpha_{1,t}(\alpha_{1,t}\alpha_{2,t}-1)(2\sigma^2+\|\beta_{1,t}\|_2^2+\|\beta_{2,t}\|_2^2)\\&\quad+\eta^2\mathbb{E}_\xi[A_t^2]+\eta^4\mathbb{E}_\xi[B_t^2]-2\eta^3\mathbb{E}_\xi[A_tB_t]+2\eta^2\mathbb{E}_\xi[B_t](\alpha_{1,t}\alpha_{2,t}-1)\\&\le(1-4\eta c)(\alpha_{1,t}\alpha_{2,t}-1)^2+2\eta\frac{1}{4}(2\sigma^2+\frac{2\eta C_2}{\sigma^2}+\frac{2\eta C_2}{\sigma^2})+\widetilde{C}_1\eta^2\\&\le(1-4\eta c)(\alpha_{1,t}\alpha_{2,t}-1)^2+\eta(\sigma^2+\frac{2\eta C_2}{\sigma^2})+\widetilde{C}_1\eta^2,\end{aligned}$$

where $\widetilde{C}_1=O\left(\frac{1}{\sigma^4}\left(\log\frac{d_1+d_2}{\delta}\right)\right)$, if $\alpha_{1,t}\alpha_{2,t}$ is of constant order. Note that here we choose $\eta\le\widetilde{\eta}_1$ as mentioned in (i). Denote $\gamma\triangleq\frac{\eta(\sigma^2+\frac{2\eta C_2}{\sigma^2})+\widetilde{C}_1\eta^2}{4\eta c}$, the inequality above can be re-expressed as

$$\mathbb{E}[\{(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2-\gamma\}|\mathcal{F}_t]\le(1-4\eta c)\{(\alpha_{1,t}\alpha_{2,t}-1)^2-\gamma\}.\tag{12}$$

We denote $G_t \triangleq (1-4\eta c)^{-t}\{(\alpha_{1,t}\alpha_{2,t}-1)^2-\gamma\}$ and $\mathcal{E}_t \triangleq \{\forall \tau \le t : (\alpha_{1,\tau}\alpha_{2,\tau}-1)^2 < (1-c)^2\}$. Since $\alpha_{1,\tau}\alpha_{2,\tau} > c$ can be inferred from $(\alpha_{1,\tau}\alpha_{2,\tau}-1)^2 < (1-c)^2$, we can get

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{E}_t}|\mathcal{F}_t] \le G_t\mathbb{1}_{\mathcal{E}_t} \le G_t\mathbb{1}_{\mathcal{E}_{t-1}}.$$

This means $\{G_t\mathbb{1}_{\mathcal{E}_{t-1}}\}$ is a supermartingale. To use Azuma's inequality, we need to bound the following difference

$$\begin{aligned} d_{t+1} &\triangleq |G_{t+1}\mathbb{1}_{\mathcal{E}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{E}_t}|\mathcal{F}_t]| \\ &= (1-4\eta c)^{-t-1}|(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2 - \mathbb{E}[(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2|\mathcal{F}_t]|\mathbb{1}_{\mathcal{E}_t} \\ &\le (1-4\eta c)^{-t-1}\widetilde{C}_2, \end{aligned}$$

where $\widetilde{C}_2 = O\left(\eta\frac{1}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)\right)$. Again, here we choose $\eta \le \widetilde{\eta}_1$. We further define $r_t \triangleq \sqrt{\sum_{i=1}^t d_i^2}$, and by Azuma's inequality we have

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathcal{E}_{t-1}} - G_0 \ge O(1)r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{O(1)r_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = O\left(\eta^2\delta\right).$$

Thus, with probability at least $1-O(\eta^2\delta)$,

$$\begin{aligned} ((\alpha_{1,t}\alpha_{2,t}-1)^2-\gamma)\mathbb{1}_{\mathcal{E}_{t-1}} &< (1-4\eta c)^t\left((\alpha_{1,0}\alpha_{2,0}-1)^2 + O(1)r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \\ &< (\alpha_{1,0}\alpha_{2,0}-1)^2 + O\left((1-4\eta c)^t r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \\ &< (1-c)^2 + (1-2c)^2 - (1-c)^2 \\ &\quad + O\left(\frac{\sqrt{\eta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right), \end{aligned}$$

When $\mathcal{E}_{t-1}$ holds, w.p. at least $1-O(\eta^2\delta)$, we will have

$$\begin{aligned} (\alpha_{1,t}\alpha_{2,t}-1)^2 &< (1-c)^2 + (1-2c)^2 - (1-c)^2 \\ &\quad + O\left(\frac{\sqrt{\eta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) + \gamma < (1-c)^2, \end{aligned}$$

as $(1-2c)^2-(1-c)^2$ is some negative constant, by choosing $\eta \le \widetilde{\eta}_2 = O\left(\sigma^4(\log\frac{1}{\delta})^{-1}\left(\log\frac{d_1+d_2}{\delta}\right)^{-1}\right)$ and σ small enough, we can make sure the sum of last four terms is negative. Now we know that if $\mathcal{E}_{t-1}$ holds, $\mathcal{E}_t$ holds w.p. at least $1-O(\eta^2\delta)$. It is easy to show that the Perturbed GD updates satisfy $(\alpha_{1,t}\alpha_{2,t}-1)^2 < (1-c)^2$ in the following $O(\frac{1}{\eta^2})$ steps w.p. at least $1-\delta$. □

3.4 Proof of Lemma 6

We partition this lemma into two parts: Lemma 5 shows that after polynomial time, the algorithm enters $\{(x,y)\big|(x^\top My-1)^2 < 4\gamma\}$, where γ is a small constant depending on σ, and Lemma 6 shows that the algorithm then stays in $\{(x,y)\big|(x^\top My-1)^2 < 6\gamma\}$. It is easy to prove Lemma 6 from Lemmas 5 and 6.

Lemma 5. Suppose $\forall t \le T_1 \triangleq O(\frac{1}{\eta^2})$,

$$(x_t,y_t) \in \left\{(x,y)|x^\top My > c\right\},$$

then $\forall\delta\in(0,1)$, with probability at least $1-\delta$, there exists a time step $\tau \le \tau_3 \triangleq O(\frac{1}{\eta}\log\frac{1}{\sigma}\log\frac{1}{\delta})$ such that

$$(x_\tau^\top My_\tau - 1)^2 < 4\gamma,$$

where $\gamma = O(\sigma^2)$.

Lemma 6. Suppose there exists a time step $\tau \le \tau_3 \triangleq O(\frac{1}{\eta}\log\frac{1}{\sigma}\log\frac{1}{\delta})$ such that

$$(x_\tau^\top M y_\tau - 1)^2 < 4\gamma,$$

and $\forall t \le T_1 \triangleq O(\frac{1}{\eta^2})$,

$$(x_t, y_t) \in \left\{(x,y) | x^\top M y > c\right\},$$

then $\forall \delta \in (0,1)$, with probability at least $1-\delta$,$\forall t \le T_1 \triangleq O(\frac{1}{\eta^2})$,

$$(x_t, y_t) \in \left\{(x,y) | (x^\top M y - 1)^2 < 6\gamma\right\}.$$

Here follows proof of Lemmas 5 and 6.

Proof. Define $\mathcal{H}_t \triangleq \{\forall \tau \le t : (\alpha_{1,\tau}\alpha_{2,\tau} - 1)^2 \ge 4\gamma\}$, for $t > \frac{\log(\frac{(\alpha_{1,0}\alpha_{2,0}-1)^2}{\gamma})}{4\eta c}$, we have

$$4\gamma\mathbb{E}[\mathbb{1}_{\mathcal{H}_t}] \le \mathbb{E}[(\alpha_{1,t}\alpha_{2,t}-1)^2 - \gamma] \le (1-4\eta c)^t\left((\alpha_{1,0}\alpha_{2,0}-1)^2 - \gamma\right) + \gamma < 2\gamma,$$

where the first inequality comes from the definition of $\mathcal{H}_t$ and the second one comes from (12). Thus, if we choose $t = O\left(\frac{\log(\frac{1}{\gamma\sigma^2})}{\eta}\right)$ and recursively applying the inequality above $O(\log(\frac{1}{\delta}))$ times, we will get, for $\tau_3 = O\left(\frac{1}{\eta}\log(\frac{1}{\delta})\log(\frac{1}{\gamma\sigma^2})\right) = O\left(\frac{1}{\eta}\log\frac{1}{\sigma}\log\frac{1}{\delta}\right)$,

$$\mathbb{P}(\mathcal{H}_{\tau_3}) < (\frac{1}{2})^{\log(\frac{1}{\delta})} = \delta.$$

Thus, w.p. at least $1-\delta$, there exists a $\tau \le \tau_3$ s.t. $(\alpha_{1,\tau}\alpha_{2,\tau}-1)^2 < 4\gamma$. Here, we finish the proof of Lemma 5.

Without loss of generality, we place the time origin at τ, i.e. $(\alpha_{1,0}\alpha_{2,0}-1)^2 < 4\gamma$. We next prove Lemma 6. Denote $\mathcal{A} \triangleq \{(\alpha_1,\alpha_2) | (\alpha_1\alpha_2 - 1)^2 < 6\gamma\}$ and $\mathcal{A}_t \triangleq \{\forall \tau \le t : (\alpha_{1,\tau}, \alpha_{2,\tau}) \in \mathcal{A}\}$, again note that $\alpha_{1,t}\alpha_{2,t} > c$ can be inferred from $(\alpha_{1,t}\alpha_{2,t}-1)^2 < 6\gamma < (1-c)^2$. Thus, without any assumption, we can get

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{A}_t}|\mathcal{F}_t] \le G_t\mathbb{1}_{\mathcal{A}_t} \le G_t\mathbb{1}_{\mathcal{A}_{t-1}}.$$

This means $\{G_t\mathbb{1}_{\mathcal{A}_{t-1}}\}$ is a supermartingale. To use Azuma's inequality, we need to bound the following difference

$$\begin{aligned}
\widetilde{d}_{t+1} &\triangleq |G_{t+1}\mathbb{1}_{\mathcal{A}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathcal{A}_t}|\mathcal{F}_t]| \\
&= (1-4\eta c)^{-t-1}|(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2 - \mathbb{E}[(\alpha_{1,t+1}\alpha_{2,t+1}-1)^2|\mathcal{F}_t]|\mathbb{1}_{\mathcal{A}_t} \\
&\le (1-4\eta c)^{-t-1}\widetilde{C}_3,
\end{aligned}$$

where $\widetilde{C}_3 = O(\sqrt{\gamma}\widetilde{C}_2)$. Further define $\widetilde{r}_t \triangleq \sqrt{\sum_{i=1}^t \widetilde{d}_i^{\,2}}$, by Azuma's inequality we have

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathcal{A}_{t-1}} - G_0 \ge O(1)\widetilde{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{O(1)\widetilde{r}_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t \widetilde{d}_i^{\,2}}\right) = O\left(\eta^2\delta\right).$$

Thus, with probability at least $1 - O(\eta^2\delta)$,

$$\begin{aligned}
((\alpha_{1,t}\alpha_{2,t}-1)^2 - \gamma)\mathbb{1}_{\mathcal{A}_{t-1}} &< (1-4\eta c)^t\left((\alpha_{1,0}\alpha_{2,0}-1)^2 + O(1)\widetilde{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \\
&< (\alpha_{1,0}\alpha_{2,0}-1)^2 + O\left((1-4\eta c)^t\widetilde{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \\
&< 4\gamma + O\left(\frac{\sqrt{\gamma\eta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}} + \sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right),
\end{aligned}$$

When $\mathcal{E}_{t-1}$ holds, w.p. at least $1-O(\eta^2\delta)$

$$(\alpha_{1,t}\alpha_{2,t}-1)^2 < 5\gamma + O\left(\frac{\widetilde{C}_3}{\sqrt{\eta}}\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) < 6\gamma,$$

by choosing $\eta \le \widetilde{\eta}_3 = O\left(\sigma^6(\log\frac{1}{\delta})^{-1}\left(\log\frac{d_1+d_2}{\delta}\right)^{-1}\right)$ we can guarantee the last term is smaller than γ. Now we know that if $\mathcal{A}_{t-1}$ holds, $\mathcal{A}_t$ holds w.p. at least $1-O(\eta^2\delta)$. It is easy to show that the Perturbed GD updates satisfy $(\alpha_{1,t}\alpha_{2,t}-1)^2 < 6\gamma$ in the following $O(\frac{1}{\eta^2})$ steps w.p. at least $1-\delta$. □

3.5 Proof of Lemma 7

Lemma 7. Suppose $(x_t^\top M y_t - 1)^2 < 6\gamma$ holds for all t, where γ is as defined above. For any $\delta \in (0,1)$ and any $\Delta > 0$, if we choose $\sigma = O\left((\log\frac{1}{\delta})^{-\frac{1}{3}}\right)$ and take step size

$$\eta \le \widetilde{\eta}_4 = O\left(\sigma^{10}\Delta\right),$$

then with probability at least $1-\delta$, we have

$$(x_t, y_t) \in \left\{(x,y)|\left((x^\top u_*)^2 - (y^\top v_*)^2\right)^2 < 6\Delta\right\},$$

for all t's such that $\tau_4 \le t \le T_1$, where $T_1 \triangleq O(\frac{1}{\eta^2})$ and $\tau_4 \triangleq O(\frac{1}{\eta\sigma^2}\log\frac{1}{\eta}\log\frac{1}{\delta})$.

Again, we partition this lemma into two parts. It is easy to prove Lemma 7 from Lemmas 8 and 9.

Lemma 8. Suppose $\forall t \le T_1 = O(\frac{1}{\eta^2})$,

$$(x_t, y_t) \in \left\{(x,y)|(x^\top M y - 1)^2 < 6\gamma\right\}.$$

then $\forall\delta \in (0,1)$, with probability at least $1-\delta$, there exists a time step $\tau \le \tau_4 = O(\frac{1}{\eta\sigma^2}\log\frac{1}{\eta}\log\frac{1}{\delta})$ such that

$$\left((x_\tau^\top u_*)^2 - (y_\tau^\top v_*)^2\right)^2 < 4\Delta,$$

where $\Delta = O(\frac{\eta}{\sigma^{10}})$.

Proof.

$$\begin{aligned}
\mathbb{E}_\xi[D_t] &= \mathbb{E}_\xi[\alpha_{1,t} < \nabla_x\mathcal{F}(x_t+\xi_{1,t}, y_t+\xi_{2,t}), u_* > -\alpha_{2,t} < \nabla_y\mathcal{F}(x_t+\xi_{1,t}, y_t+\xi_{2,t}), v_* >] \\
&= \alpha_{1,t}(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|_2^2) - \alpha_{2,t} + \alpha_{1,t}\sigma^2) - \alpha_{2,t}(\alpha_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|_2^2) - \alpha_{1,t} + \alpha_{2,t}\sigma^2) \\
&= \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2) + (\alpha_{1,t}^2\|\beta_{2,t}\|_2^2 - \alpha_{2,t}^2\|\beta_{1,t}\|_2^2).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}_\xi[D_t](\alpha_{1,t}^2 - \alpha_{2,t}^2) &= \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 + (\alpha_{1,t}^2 - \alpha_{2,t}^2)(\alpha_{1,t}^2\|\beta_{2,t}\|_2^2 - \alpha_{2,t}^2\|\beta_{1,t}\|_2^2) \\
&\ge \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 2\alpha_{1,t}^2\alpha_{2,t}^2\frac{2\eta C_2}{\sigma^2} \\
&> \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 2(1+\sqrt{6\gamma})^2\frac{2\eta C_2}{\sigma^2} \\
&> \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 3\frac{2\eta C_2}{\sigma^2}.
\end{aligned}$$

The last inequality can be achieved by choosing $\gamma < \frac{(\sqrt{3/2}-1)^2}{6}$. Plugging (5) and (6) into (4), taking expectation conditioning on previous trajectory $\mathcal{F}_t$ and plugging the equation above in, we get

$$\begin{aligned}
\mathbb{E}[(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2|\mathcal{F}_t] &= (\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 4\eta\mathbb{E}_\xi[D_t](\alpha_{1,t}^2 - \alpha_{2,t}^2) \\
&\quad + 4\eta^2\mathbb{E}_\xi[D_t^2] + \eta^4\mathbb{E}_\xi[F_t^2] - 4\eta^3\mathbb{E}_\xi[D_tF_t] + 2\eta^2\mathbb{E}_\xi[F_t]\left(\alpha_{1,t}^2 - \alpha_{2,t}^2\right) \\
&\le \left(1 - 4\eta\sigma^2\right)(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 + 24\eta\frac{\eta C_2}{\sigma^2} + \widetilde{C}_4\eta^2,
\end{aligned}$$

where $\widetilde{C}_4 = O(\frac{1}{\sigma^4}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)^2)$. Denote $\Delta \triangleq \frac{24\eta\frac{\eta C_2}{\sigma^2}+\widetilde{C}_4\eta^2}{4\eta\sigma^2}$. the inequality above can be re-expressed as

$$\mathbb{E}[\{(\alpha_{1,t+1}^2-\alpha_{2,t+1}^2)^2-\Delta\}|\mathcal{F}_t] \le \left(1-4\eta\sigma^2\right)\{(\alpha_{1,t}^2-\alpha_{2,t}^2)^2-\Delta\}.$$

Denote $\mathcal{B}_t \triangleq \{\forall\tau\le t:((\alpha_{1,\tau}^2-\alpha_{2,\tau}^2)^2\ge 4\Delta\}$, for $t > \frac{\log(\frac{\left(\alpha_{1,0}^2-\alpha_{2,0}^2\right)^2}{\frac{\delta}{4\eta\sigma^2}})}{4\eta\sigma^2}$, we have

$$4\Delta\mathbb{E}[\mathbb{1}_{\mathcal{B}_t}] \le \mathbb{E}[(\alpha_{1,t}^2-\alpha_{2,t}^2)^2-\Delta] \le (1-4\eta c)^t((\alpha_{1,0}^2-\alpha_{2,0}^2)^2-\Delta)+\Delta < 2\Delta,$$

where the first inequality comes from the definition of $\mathcal{B}_t$ and the second one comes from calculation above. Thus, if we choose $t = O(\frac{\log(\frac{1}{\Delta\sigma^2})}{\eta\sigma^2})$ and recursively applying the inequality above $\log(\frac{1}{\delta})$ times, we will get, for $\tau_4 = O(\frac{1}{\eta\sigma^2}\log(\frac{1}{\delta})\log(\frac{1}{\Delta\sigma^2})) = O(\frac{1}{\eta\sigma^2}\log(\frac{1}{\delta})\log(\frac{1}{\eta}))$,

$$\mathbb{P}(\mathcal{B}_{\tau_4}) < (\frac{1}{2})^{\log(\frac{1}{\delta})} = \delta.$$

Thus, w.p. at least $1-\delta$, there exists a $\tau\le\tau_4$ s.t. $(\alpha_{1,\tau}^2-\alpha_{2,\tau}^2)^2 < 4\Delta$. Here, we finish the proof of Lemma 8. □

Lemma 9. Suppose there exists a time step $\tau\le\tau_4 = O(\frac{1}{\eta\sigma^2}\log\frac{1}{\Delta\sigma^2}\log\frac{1}{\delta})$ such that

$$\left((x_\tau^\top u_*)^2-(y_\tau^\top v_*)^2\right)^2 < 4\Delta,$$

then $\forall\delta\in(0,1)$, with probability at least $1-\delta$, $\forall t\le T_1 \triangleq O(\frac{1}{\eta^2})$,

$$(x_t,y_t)\in\left\{(x,y)|\left((x^\top u_*)^2-(y^\top v_*)^2\right)^2 < 6\Delta\right\}.$$

Proof. Without loss of generality, we place the time origin at τ, i.e. $(\alpha_{1,0}^2-\alpha_{2,0}^2)^2 < 4\Delta$. Denote $\mathcal{D} \triangleq \{(\alpha_1,\alpha_2)|(\alpha_1^2-\alpha_2^2)^2<6\Delta\}$ and $\mathcal{D}_t \triangleq \{\forall\tau\le t:(\alpha_{1,\tau},\alpha_{2,\tau})\in\mathcal{D}\}$. Note that $(\alpha_{1,t}\alpha_{2,t}-1)^2<6\gamma$ still holds with high probability. Defining $H_t \triangleq (1-4\eta\sigma^2)^{-t}\{(\alpha_{1,t}^2-\alpha_{2,t}^2)^2-\Delta\}$, we can get

$$\mathbb{E}[H_{t+1}\mathbb{1}_{\mathcal{D}_t}|\mathcal{F}_t] \le H_t\mathbb{1}_{\mathcal{D}_t} \le H_t\mathbb{1}_{\mathcal{D}_{t-1}}.$$

This means $\{H_t\mathbb{1}_{\mathcal{D}_{t-1}}\}$ is a supermartingale. To use Azuma's inequality, we need to bound the following difference

$$\begin{aligned}\bar{d}_{t+1} &\triangleq |H_{t+1}\mathbb{1}_{\mathcal{D}_t}-\mathbb{E}[H_{t+1}\mathbb{1}_{\mathcal{D}_t}|\mathcal{F}_t]|\\ &= (1-4\eta\sigma^2)^{-t-1}|(\alpha_{1,t+1}^2-\alpha_{2,t+1}^2)^2-\mathbb{E}[(\alpha_{1,t+1}^2-\alpha_{2,t+1}^2)^2|\mathcal{F}_t]|\mathbb{1}_{\mathcal{D}_t}\\ &\le (1-4\eta\sigma^2)^{-t-1}\widetilde{C}_5,\end{aligned}$$

where $\widetilde{C}_5 = O(\eta\frac{\sqrt{\Delta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right))$. Further define $\bar{r}_t \triangleq \sqrt{\sum_{i=1}^t\bar{d}_i^{\,2}}$, by Azuma's inequality we will get

$$\mathbb{P}\left(H_t\mathbb{1}_{\mathcal{D}_{t-1}}-H_0\ge O(1)\bar{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right)\le\exp\left(-\frac{O(1)\bar{r}_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t\bar{d}_i^{\,2}}\right)=O\left(\eta^2\delta\right).$$

Thus, with probability at least $1-O(\eta^2\delta)$, we have

$$\begin{aligned}((\alpha_{1,t}^2-\alpha_{2,t}^2)^2-\Delta)\mathbb{1}_{\mathcal{D}_{t-1}} &< (1-4\eta\sigma^2)^t\left((\alpha_{1,0}^2-\alpha_{2,0}^2)^2+O(1)\bar{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right)\\ &< (\alpha_{1,0}^2-\alpha_{2,0}^2)^2+O\left((1-4\eta\sigma^2)^t\bar{r}_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right)\\ &< 4\Delta+O\left(\frac{\sqrt{\Delta\eta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}}+\sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right).\end{aligned}$$

When $\mathcal{D}_{t-1}$ holds, w.p. at least $1 - O(\eta^2\delta)$, by the inequality above we have

$$(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 < 5\Delta + O\left(\frac{\sqrt{\Delta\eta}}{\sigma^2}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}} + \sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) < 6\Delta.$$

Note that to make sure last terms is smaller than Δ, we need

$$\frac{\sqrt{\eta}}{\sigma^2\sqrt{\Delta}}\left(\sqrt{\log\frac{d_1+d_2}{\eta^2}} + \sqrt{\log\frac{1}{\delta}}\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) = O(1).$$

As $\Delta = O(\frac{\eta}{\sigma^{10}})$, we know that as long as η is polynomial in σ, choosing $\sigma = O((\log\frac{1}{\delta})^{-\frac{1}{3}})$ is sufficient. Now we know that if $\mathcal{D}_{t-1}$ holds, $\mathcal{D}_t$ holds w.p. at least $1 - O(\eta^2\delta)$. It is easy to show that the Perturbed GD updates satisfy $(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 < 6\Delta$ in the following $O(\frac{1}{\eta^2})$ steps w.p. at least $1 - \delta$. □

With Lemmas 1, 6 and 7, we can prove Lemma 7. Here follows a brief proof.

Proof.

$$\begin{aligned}
|1 - x_t^\top u_*| &< (1 + x_t^\top u_*)|1 - x_t^\top u_*| \\
&= |1 - (x_t^\top u_*)^2 + x_t^\top u_* v_*^\top y_t - x_t^\top u_* v_*^\top y_t| \\
&\le |1 - x_t^\top u_* v_*^\top y_t| + |(x_t^\top u_*)^2 - x_t^\top u_* v_*^\top y_t| \\
&= |1 - x_t^\top M y_t| + x_t^\top u_* |x_t^\top u_* - v_*^\top y_t| \\
&\le |1 - x_t^\top M y_t| + \frac{\sqrt{2}}{\sigma}|x_t^\top u_* - v_*^\top y_t| \\
&< \sqrt{6\gamma} + \frac{\sqrt{2}}{\sigma}\sqrt{6\Delta}.
\end{aligned}$$

The last inequality comes from Lemmas 6 and 9. Together with Lemma 1 we can get

$$\|x_t - u_*\|^2 = (1 - x_t^\top u_*)^2 + \|\beta_{1,t}\|_2^2 < (\sqrt{6\gamma} + \frac{\sqrt{2}}{\sigma}\sqrt{6\Delta})^2 + 2\frac{\eta C_2}{\sigma^2} = O(\sigma^2 + \frac{\eta}{\sigma^{10}}).$$

Note that we use $C_2 = O(\frac{1}{\sigma^6})$ when calculating the order. Similarly, we have

$$\|y_t - v_*\|^2 < (\sqrt{6\gamma} + \frac{\sqrt{2}}{\sigma}\sqrt{6\Delta})^2 + 2\frac{\eta C_2}{\sigma^2} = O(\sigma^2 + \frac{\eta}{\sigma^{10}}).$$

Then, for any $\epsilon > 0$, by choosing $\sigma = O(\sqrt{\epsilon})$ and $\eta \le \eta_5 = O(\sigma^{10}\epsilon)$, we will have $\|x_t - u_*\|^2 < \epsilon$ and $\|y_t - v_*\|^2 < \epsilon$ □

4 Proof of Theorem 2

Recall that the gradient of $\widetilde{\mathcal{F}}$ takes the following form.

$$\begin{aligned}
\nabla_X \widetilde{\mathcal{F}}(X,Y) &= (XY^\top - M)Y - d_2\sigma_2^2 X, \\
\nabla_Y \widetilde{\mathcal{F}}(X,Y) &= (XY^\top - M)^\top X - d_1\sigma_1^2 Y.
\end{aligned}$$

Suppose (U,V) is a stationary point. Then we have

$$(UV^\top - M)V - d_2\sigma_2^2 U = 0, \tag{13}$$
$$(UV^\top - M)^\top U - d_1\sigma_1^2 V = 0. \tag{14}$$

• **Step 1:** To prove the first statement, simply left multiply each side of (13) by $U^\top$ and each side of (14) by $V^\top$, and we have the following equations.

$$\begin{aligned}
U^\top UV^\top V - U^\top MV - d_2\sigma_2^2 U^\top U = 0, \\
V^\top VU^\top U - V^\top M^\top U - d_1\sigma_1^2 V^\top V = 0.
\end{aligned}$$

Note that the following equation naturally holds.

$$U^\top UV^\top V - U^\top MV = \left(V^\top VU^\top U - V^\top M^\top U\right)^\top.$$

Combine these three equations together and we have

$$U^\top U = \frac{d_1\sigma_1^2}{d_2\sigma_2^2}(V^\top V)^\top = \gamma^2 V^\top V.$$

• **Step 2:** We next show that $(\widetilde{U}, \widetilde{V})R$ is a stationary point, where

$$(\widetilde{U}, \widetilde{V}) = (\sqrt{\gamma}A(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}, \frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}),$$

where $R \in \mathbb{R}^{r\times r}$ is an orthogonal matrix. We only need to check (13) and (14). In fact, we have

$$\begin{aligned}\nabla_X \widetilde{\mathcal{F}}(\widetilde{U}R, \widetilde{V}R) &= -\gamma\sigma^2 AB^\top \frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2\sqrt{\gamma}A(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R \\ &= -\sigma^2\sqrt{\gamma}A(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2\sqrt{\gamma}A(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R \\ &= 0,\end{aligned}$$

and

$$\begin{aligned}\nabla_Y \widetilde{\mathcal{F}}(\widetilde{U}R, \widetilde{V}R) &= -\gamma\sigma^2 BA^\top \sqrt{\gamma}A(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2\gamma^2\frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R \\ &= -\sigma^2\gamma\sqrt{\gamma}B(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2\gamma\sqrt{\gamma}B(\Sigma - \gamma\sigma^2 I_r)^{\frac{1}{2}}R \\ &= 0.\end{aligned}$$

Combine the above equations together and we know that $(\widetilde{U}, \widetilde{V})$ is a stationary point.

• **Step 3:** We next show that $\{(\widetilde{U}, \widetilde{V})R \big| R \in \mathbb{R}^{r\times r}, \text{orthogonal}\}$ are the global minima and all other stationary points enjoy strict saddle property. Without loss of generality, we assume $\gamma = 1$.

We first calculate the Hessian $\nabla^2\widetilde{\mathcal{F}}(X, Y)$. The Hessian can be viewed as a matrix that operates on vectorized matrices of dimension $(d_1 + d_2) \times r$. Then, for any $W \in \mathbb{R}^{(d_1+d_2)\times r}$, the Hessian defines a quadratic form

$$[\nabla^2\widetilde{\mathcal{F}}(W)](Z_1, Z_2) = \sum_{i,j,k,l}\frac{\partial^2\widetilde{\mathcal{F}}(W)}{\partial W[i,j]\partial W[k,l]}Z_1[i,j]Z_2[k,l],\ \forall Z_1, Z_2 \in \mathbb{R}^{(d_1+d_2)\times r}.$$

We can then express the Hessian $\nabla^2\widetilde{\mathcal{F}}(W)$ as follows:

$$[\nabla^2\widetilde{\mathcal{F}}(X, Y)](\Delta, \Delta) = 2 < XY^\top - M, \Delta_U\Delta_V^\top > + \|U\Delta_V^\top + \Delta_U V^\top\|_{\mathrm{F}}^2 + \sigma^2\|\Delta_U\|_{\mathrm{F}}^2 + \sigma^2\|\Delta_V\|_{\mathrm{F}}^2,$$

where $\Delta = \begin{bmatrix}\Delta_U \\ \Delta_V\end{bmatrix}$, $\Delta_U \in \mathbb{R}^{d_1\times r}$ and $\Delta_V \in \mathbb{R}^{d_2\times r}$. We further denote $W = \begin{bmatrix}X \\ Y\end{bmatrix}$, $\widetilde{W} = \begin{bmatrix}\widetilde{U} \\ \widetilde{V}\end{bmatrix}$, and $\widetilde{M} = \widetilde{U}\widetilde{V}^\top$,

$$R = \underset{R'\in\mathbb{R}^{r\times r}, \text{orthogonal}}{\operatorname{argmin}} \|W - \widetilde{W}R'\|.$$

We then have the following lemma.

Lemma 10. Let $\sigma_{\min}(M)$ be the smallest singular value of M. Suppose $d_1\sigma_1^2 = d_2\sigma_2^2 = \sigma^2$, and $\sigma^2 < \sigma_{\min}(M)$. For $\forall(U, V) \in (\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ such that $\nabla\widetilde{\mathcal{F}}(U, V) = 0$, we denote $\Delta = \begin{bmatrix}U - \widetilde{U}R \\ V - \widetilde{V}R\end{bmatrix}$, then we have

$$[\nabla^2\widetilde{\mathcal{F}}(U, V)](\Delta, \Delta) \le -\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 - 3\sigma^2\|A^\top U - B^\top V\|_{\mathrm{F}}^2. \tag{15}$$

Moreover, $[\nabla^2\widetilde{\mathcal{F}}(U, V)](\Delta, \Delta) < 0$ if

$$(U, V) \notin \{(\widetilde{U}, \widetilde{V})R' \big| R' \in \mathbb{R}^{r\times r}, R'R'^\top = R'^\top R' = I_r\}.$$

Proof. Recall that the quadratic form defined by the Hessian can be written as follows.

$$[\nabla^2 \widetilde{\mathcal{F}}(U,V)](\Delta,\Delta) = 2 < UV^\top - M, \Delta_U \Delta_V^\top > + \|U\Delta_V^\top + \Delta_U V^\top\|_{\mathrm{F}}^2 + \sigma^2\|\Delta_U\|_{\mathrm{F}}^2 + \sigma^2\|\Delta_V\|_{\mathrm{F}}^2. \tag{16}$$

We start from the second term $\|U\Delta_V^\top + \Delta_U V^\top\|_{\mathrm{F}}^2$. Similar to the proof of Claim B.5 in Du et al. (2018), we have

$$\begin{aligned}
\|U\Delta_V^\top + \Delta_U V^\top\|_{\mathrm{F}}^2 &= \|\Delta_U\Delta_V^\top + UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 \\
&= \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + 2 < \Delta_U\Delta_V^\top, UV^\top - \widetilde{M} > \\
&= \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + 2 < \Delta_U\Delta_V^\top, UV^\top - M > + 2 < \Delta_U\Delta_V^\top, M - \widetilde{M} > \\
&= \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + 2 < \Delta_U\Delta_V^\top, UV^\top - M > + 2\sigma^2 < \Delta_U\Delta_V^\top, AB^\top >
\end{aligned}$$

Plugging this equation into (16), we have

$$\begin{aligned}
[\nabla^2 \widetilde{\mathcal{F}}(U,V)](\Delta,\Delta) = 4 < UV^\top - M, \Delta_U\Delta_V^\top > &+ \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 \\
&+ \sigma^2\left(\|\Delta_U\|_{\mathrm{F}}^2 + \|\Delta_V\|_{\mathrm{F}}^2 + 2 < \Delta_U\Delta_V^\top, AB^\top >\right).
\end{aligned}$$

Note that using the fact $\nabla\widetilde{\mathcal{F}}(U,V) = \nabla\widetilde{\mathcal{F}}(\widetilde{U},\widetilde{V}) = 0$, one can easily verify that

$$\begin{aligned}
4 < UV^\top - M, \Delta_U\Delta_V^\top > &= 4 < UV^\top - M, \widetilde{M} > -2\sigma^2\left(\|\Delta_U\|_{\mathrm{F}}^2 + \|\Delta_V\|_{\mathrm{F}}^2\right) + 2\sigma^2\left(\|\widetilde{U}\|_{\mathrm{F}}^2 + \|\widetilde{V}\|_{\mathrm{F}}^2\right) \\
&= -4\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 - 4\sigma^2 < AB^\top, \widetilde{M} - UV^\top > \\
&\quad + 2\sigma^2\left(\|\widetilde{U}\|_{\mathrm{F}}^2 + \|\widetilde{V}\|_{\mathrm{F}}^2 - \|U\|_{\mathrm{F}}^2 - \|V\|_{\mathrm{F}}^2\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
[\nabla^2 \widetilde{\mathcal{F}}(U,V)](\Delta,\Delta) = -4\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 &+ \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 \\
&- \sigma^2\Bigg[\|\Delta_U\|_{\mathrm{F}}^2 + \|\Delta_V\|_{\mathrm{F}}^2 - 2 < \Delta_U\Delta_V^\top, AB^\top > \\
&\quad - 2\left(\|\widetilde{U}\|_{\mathrm{F}}^2 + \|\widetilde{V}\|_{\mathrm{F}}^2 - \|U\|_{\mathrm{F}}^2 - \|V\|_{\mathrm{F}}^2\right) + 4 < AB^\top, \widetilde{M} - UV^\top >\Bigg].
\end{aligned} \tag{17}$$

We then have the following two claims:

Claim 1. $-4\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 \le -\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2$.

Proof. Similar to the proof of Claim B.5 in Du et al. (2018), we have

$$\begin{aligned}
\|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 &\le \frac{1}{4}\|\Delta\Delta^\top\|_{\mathrm{F}}^2 \\
&\le \frac{1}{2}\|WW^\top - \widetilde{W}\widetilde{W}^\top\|_{\mathrm{F}}^2 \\
&= 2\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 - \|U^\top\widetilde{U} - V^\top\widetilde{V}\|_{\mathrm{F}}^2 + \frac{1}{2}\|U^\top U - V^\top V\|_{\mathrm{F}}^2 + \frac{1}{2}\|\widetilde{U}^\top\widetilde{U} - \widetilde{V}^\top\widetilde{V}\|_{\mathrm{F}}^2 \\
&= 2\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 - \|U^\top\widetilde{U} - V^\top\widetilde{V}\|_{\mathrm{F}}^2 \\
&\le 2\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
-4\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + \|\Delta_U\Delta_V^\top\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 &\le -4\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + 2\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 + \|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2 \\
&= -\|UV^\top - \widetilde{M}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

Claim 2. $\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2 - \|\widetilde{U}\|_{\mathrm{F}}^2 - \|\widetilde{V}\|_{\mathrm{F}}^2 + 2 < AB^\top, \widetilde{M} - UV^\top >= \|A^\top\Delta_U - B^\top\Delta_V\|_{\mathrm{F}}^2$.

Proof. First, the LHS of the equation can be rewritten as follows.

$$\begin{aligned}
&\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2-\|\widetilde{U}\|_{\mathrm{F}}^2-\|\widetilde{V}\|_{\mathrm{F}}^2+2<AB^\top,\widetilde{M}-UV^\top>\\
=&\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2-2<AB^\top,\widetilde{M}+UV^\top>+4<AB^\top,\widetilde{M}>-\|\widetilde{U}\|_{\mathrm{F}}^2-\|\widetilde{V}\|_{\mathrm{F}}^2\\
=&\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2-2<AB^\top,\widetilde{M}+UV^\top>+4\|\widetilde{U}\|_{\mathrm{F}}^2-\|\widetilde{U}\|_{\mathrm{F}}^2-\|\widetilde{V}\|_{\mathrm{F}}^2\\
=&\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2+\|\widetilde{U}\|_{\mathrm{F}}^2+\|\widetilde{V}\|_{\mathrm{F}}^2-2<AB^\top,\widetilde{M}+UV^\top>,
\end{aligned}$$

where we use the fact

$$<AB^\top,\widetilde{M}>=\|\widetilde{U}\|_{\mathrm{F}}^2=\|\widetilde{V}\|_{\mathrm{F}}^2.$$

On the other hand, the RHS of the equation can be rewritten as follows.

$$\begin{aligned}
\|A^\top\Delta_U-B^\top\Delta_V\|_{\mathrm{F}}^2&=\|\Delta_U\|_{\mathrm{F}}^2+\|\Delta_V\|_{\mathrm{F}}^2-2<\Delta_U\Delta_V^\top,AB^\top>\\
&=\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2+\|\widetilde{U}\|_{\mathrm{F}}^2+\|\widetilde{V}\|_{\mathrm{F}}^2-2<AB^\top,\widetilde{M}+UV^\top>\\
&\quad-2\operatorname{tr}(U(\widetilde{U}R)^\top)-2\operatorname{tr}(V(\widetilde{V}R)^\top)+2\operatorname{tr}(U(\widetilde{U}R)^\top)+2\operatorname{tr}(V(\widetilde{V}R)^\top)\\
&=\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2+\|\widetilde{U}\|_{\mathrm{F}}^2+\|\widetilde{V}\|_{\mathrm{F}}^2-2<AB^\top,\widetilde{M}+UV^\top>\\
&=\|U\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2-\|\widetilde{U}\|_{\mathrm{F}}^2-\|\widetilde{V}\|_{\mathrm{F}}^2+2<AB^\top,\widetilde{M}-UV^\top>.
\end{aligned}$$

□

Plugging the conclusions in Claims 1 and 2 into (17), we have

$$[\nabla^2\widetilde{\mathcal{F}}(U,V)](\Delta,\Delta)\le-\|UV^\top-\widetilde{M}\|_{\mathrm{F}}^2-3\sigma^2\|A^\top\Delta_U-B^\top\Delta_V\|_{\mathrm{F}}^2.$$

Note that since $A^\top\widetilde{U}R=B^\top\widetilde{V}R$, we have $A^\top\Delta_U-B^\top\Delta_V=A^\top U-B^\top V$. To justify our last statement, we have the following claim.

Claim 3. $\|UV^\top-\widetilde{M}\|_{\mathrm{F}}^2+3\sigma^2\|A^\top U-B^\top V\|_{\mathrm{F}}^2=0$ if and only if

$$(U,V)=(\widetilde{U},\widetilde{V})R,$$

where $R\in\mathbb{R}^{r\times r}$ is an orthogonal matrix.

Proof. $\|UV^\top-\widetilde{M}\|_{\mathrm{F}}^2+3\sigma^2\|A^\top U-B^\top V\|_{\mathrm{F}}^2=0$ if and only if

$$UV^\top=\widetilde{M},\tag{18}$$
$$A^\top U=B^\top V.\tag{19}$$

Left multiplying each side of (18) by $A^\top$, we have

$$\begin{aligned}
A^\top UV^\top=(\Sigma-\sigma^2I)B^\top&\Leftrightarrow B^\top VV^\top=(\Sigma-\sigma^2I)B^\top\\
&\Leftrightarrow VV^\top=B(\Sigma-\sigma^2I)B^\top\\
&\Leftrightarrow V=B(\Sigma-\sigma^2I)^{\frac{1}{2}}R=\widetilde{V}R'.
\end{aligned}$$

The last equivalent argument comes from Theorem 4 in Li et al. (2019). Plugging $V=\widetilde{V}R'$ back to (19), we have $U=\widetilde{U}R'$. □

As a direct result of Claim 3, for stationary point $(U,V)\notin\{(\widetilde{U},\widetilde{V})R'\big|R'\in\mathbb{R}^{r\times r},R'R'^\top=R'^\top R'=I_r\}$, we have

$$[\nabla^2\widetilde{\mathcal{F}}(U,V)](\Delta,\Delta)<0.$$

We prove the lemma. □

Lemma 10 directly implies that $\{(\widetilde{U},\widetilde{V})R'\big|R'\in\mathbb{R}^{r\times r},R'R'^\top=R'^\top R'=I_r\}$ contains all the global optima, and all other stationary points enjoy strict saddle property.

5 Perturbed GD

The detail of the Perturbed GD algorithm is summarized in Algorithm 1.

Algorithm 1 Perturbed Gradient Descent for Rank-1 Matrix Factorization.

Input: step size η, noise level σ_1, σ_2, matrix $M \in \mathbb{R}^{d_1 \times d_2}$, number of iterations T.
Initialize: initialize (x_0, y_0) arbitrarily.
for $t = 0 \dots T-1$ **do**
  Sample $\xi_{1,t} \sim N(0, \sigma_1^2 I_{d_1})$ and $\xi_{2,t} \sim N(0, \sigma_2^2 I_{d_2})$.
  $\widetilde{x}_t = x_t + \xi_{1,t},\ \widetilde{y}_t = y_t + \xi_{2,t}$.
  $x_{t+1} = x_t - \eta(\widetilde{x}_t \widetilde{y}_t^\top - M)\widetilde{y}_t$.
  $y_{t+1} = y_t - \eta(\widetilde{y}_t \widetilde{x}_t^\top - M^\top)\widetilde{x}_t$.
end for

6 Figures

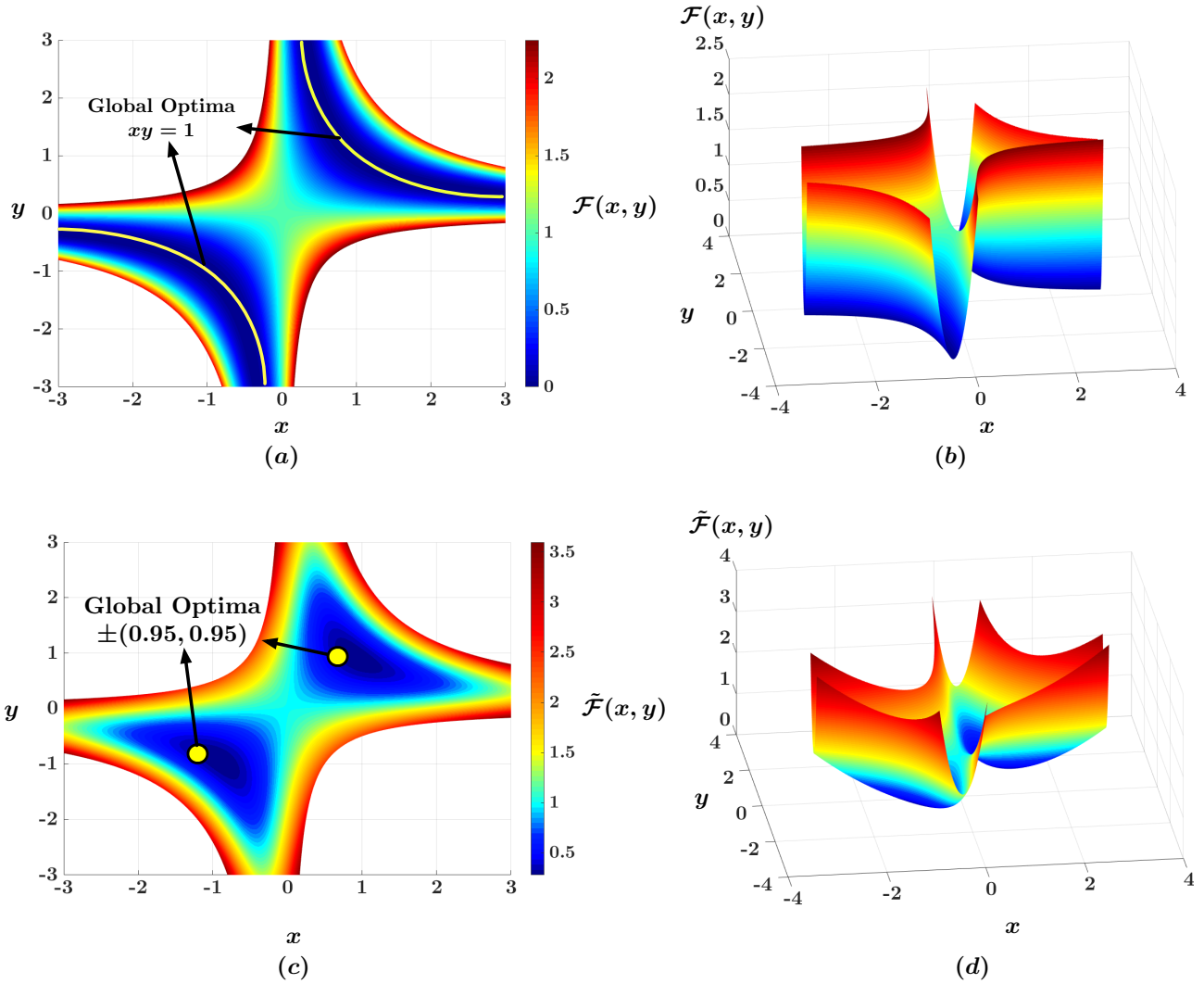

Figure 1: The visualization of objective functions $\mathcal{F}(x, y) = (1 - xy)^2$ and $\widetilde{\mathcal{F}}(x, y)$ with $x, y \in \mathbb{R}$ and $\sigma_1^2 = \sigma_2^2 = 0.0975$. For $\mathcal{F}(x, y)$, any (x, y) that satisfies $xy = 1$ is a global minimum (as shown in (a)). $\widetilde{\mathcal{F}}(x, y)$ only has global minima close to $\pm(1, 1)$ (as shown in (c)).

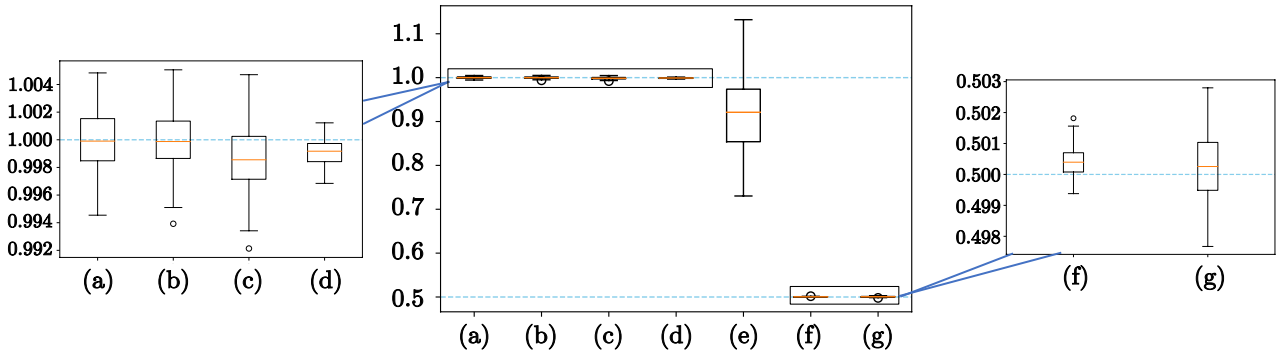

Figure 2: P-GD with balanced noise (a, b and c), P-GD with unbalanced noise (g) and GD (d,e and f) for the rank-1 matrix factorization problem. (a) and (d) use small initializations ($\sigma_x = \sigma_y = 10^{-2}$) and balanced step size ($\eta_x = \eta_y = 10^{-2}$). (b) and (e) use large initializations ($\sigma_x = \sigma_y = 10^{-1}$) and balanced step size. (c) and (f) use small initializations and unbalanced step size ($\eta_x = 0.5\eta_y = 5 \times 10^{-3}$).

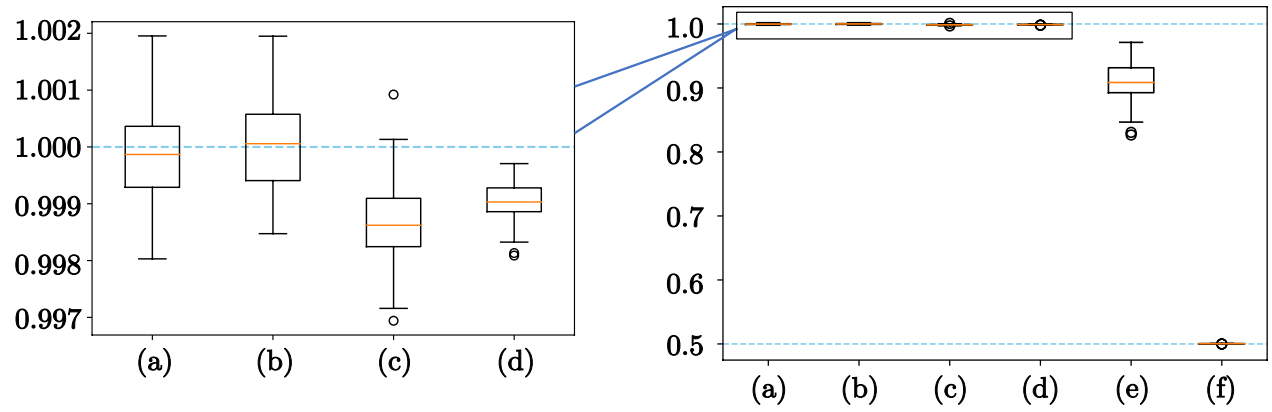

Figure 3: P-GD with balanced noise (a, b and c) and GD (d,e and f) for rank-10 matrix factorization problem. (a) and (d) use small initializations ($\sigma_x = \sigma_y = 10^{-2}$) and balanced step size ($\eta_x = \eta_y = 10^{-2}$). (b) and (e) use large initializations ($\sigma_x = \sigma_y = 10^{-1}$) and balanced step size. (c) and (f) use small initializations and unbalanced step size ($\eta_x = 0.5\eta_y = 5 \times 10^{-3}$).

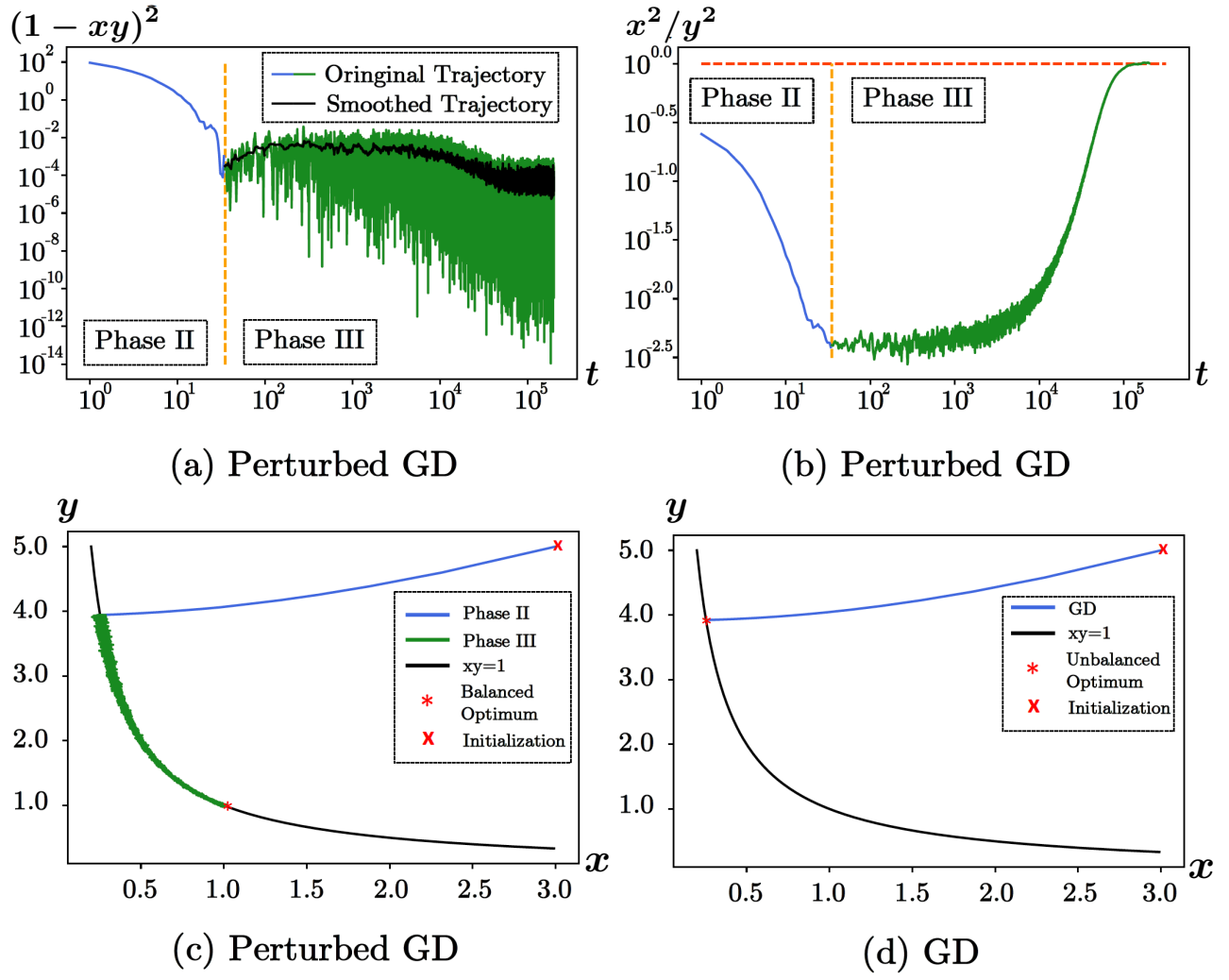

Figure 4: Algorithmic behaviors of P-GD and GD. For P-GD, phase transition happens around the first $30\sim40$ iterations, as shown in (a,b,c). GD does not show phase transitions.

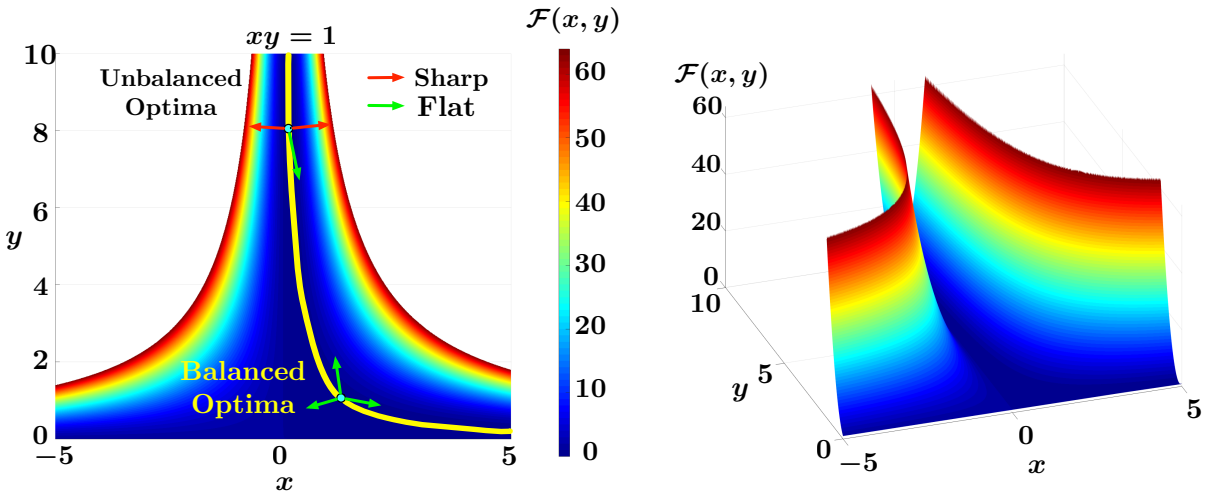

Figure 5: The visualization of objective $\mathcal{F}(x, y) = (1 - xy)^2$. All the global optima are connected and form a path. The landscape around the path forms a valley. Around unbalanced optima, the landscape is sharp in some directions and flat in others. Around balanced optima, the landscape only contains flat directions.

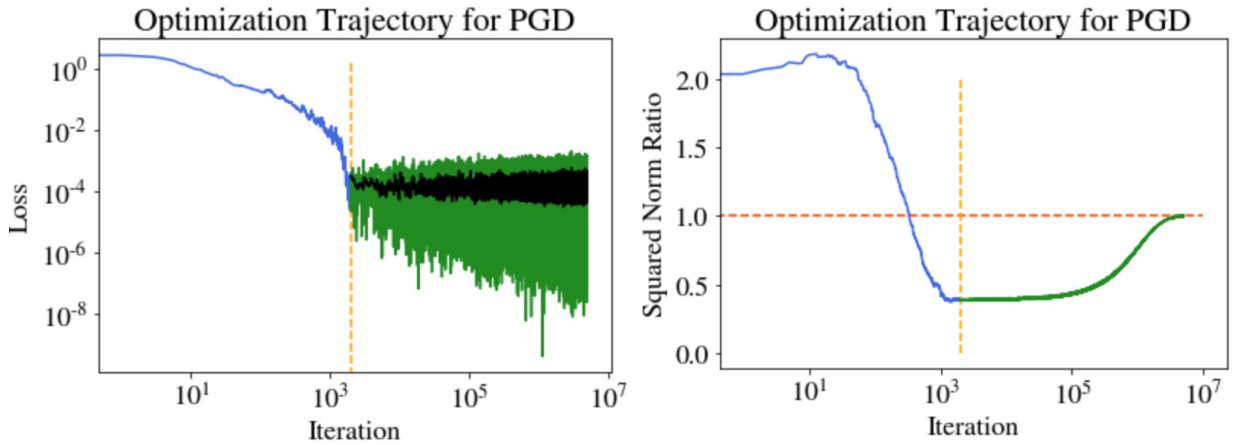

Figure 6: Algorithmic behaviors of Perturbed GD and GD for $d = 4$. For Perturbed GD, phase transition happens around the first 2×10^3 iterations.

References

Du, S. S., Hu, W., and Lee, J. D. (2018). Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395.

Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., and Zhao, T. (2019). Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514.