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 \( \{u_*, \bar{u}_1, \ldots, \bar{u}_{d_1-1}\} \) and \( \{v_*, \bar{v}_1, \ldots, \bar{v}_{d_2-1}\} \) 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

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

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

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

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

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

\[
x_{t+1}^T My_{t+1} = \alpha_{1,t+1} \alpha_{2,t+1} \\
= (\alpha_{1,t} - \eta < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_*>)(\alpha_{2,t} - \eta < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_*>) \\
= \alpha_{1,t} \alpha_{2,t} - \eta (\alpha_{2,t} < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_*> + \alpha_{1,t} < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_*>) \\
+ \eta^2 < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_*> < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_*> . \quad (1)
\]

For simplicity, we denote

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

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

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

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

\[
(x_{t+1}^T My_{t+1} - 1)^2 = (\alpha_{1,t+1} \alpha_{2,t+1} - 1)^2 \\
= (\alpha_{1,t} \alpha_{2,t} - 1 - \eta A_t + \eta^2 B_t)^2 \\
= (\alpha_{1,t} \alpha_{2,t} - 1)^2 + \eta^2 A_t^2 + \eta^4 B_t^2 \\
- 2\eta A_t (\alpha_{1,t} \alpha_{2,t} - 1) - 2\eta^3 A_t B_t + 2\eta^2 B_t (\alpha_{1,t} \alpha_{2,t} - 1) . \quad (3)
\]
Furthermore, since the balanced optima satisfy $x^T u_* = y^T v_*$, we further explicitly write down the update of $(x^T u_*)^2 - (y^T v_*)^2$ as follows.

$$
((x_{t+1}^T u_*)^2 - (y_{t+1}^T v_*)^2)^2 = (\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2
= (\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 + 4\eta^2 D_t^2 + 4\eta^4 F_t^2
- 4\eta D_t (\alpha_{1,t}^2 - \alpha_{2,t}^2) - 4\eta^3 D_t F_t + 2\eta^2 F_t (\alpha_{1,t}^2 - \alpha_{2,t}^2),
$$

(4)

where $D_t$ and $F_t$ is defined as

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

$$
F_t \triangleq < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >^2 - < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >^2.
$$

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

$$
< \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_* >
= u_*^T \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t})
= u_*^T (x_t + \xi_{1,t})(y_t + \xi_{2,t})^T - u_* v_*^T (y_t + \xi_{2,t})
= \alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2) - \alpha_{2,t} + 2\alpha_{1,t} y_*^T \xi_{2,t} - v_*^T \xi_{2,t} + u_*^T \xi_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2)

+ 2u_*^T \xi_{1,t} y_*^T \xi_{2,t} + \alpha_{1,t}\|\xi_{2,t}\|^2 + u_*^T \xi_{1,t}\|\xi_{2,t}\|^2
\triangleq \alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2) - \alpha_{2,t} + g_x,
$$

(5)

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

Similarly, we have

$$
< \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_* >
= \alpha_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|^2) - \alpha_{1,t} + 2\alpha_{2,t} x_*^T \xi_{1,t} - u_*^T \xi_{1,t} + v_*^T \xi_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|^2)

+ 2v_*^T \xi_{2,t} x_*^T \xi_{1,t} + \alpha_{2,t}\|\xi_{1,t}\|^2 + v_*^T \xi_{2,t}\|\xi_{1,t}\|^2
\triangleq \alpha_{2,t}(\alpha_{1,t}^2 + \|\beta_{1,t}\|^2) - \alpha_{1,t} + g_y,
$$

(6)

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

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

2.1 Proof of Lemma 1

Proof. By setting the gradient of $F$ to zero we get

$$
\|x\|^2 y = M^T x,
$$

(7)

$$
\|y\|^2 x = M y.
$$

(8)

Recall that $M = u_* v_*^T$ and

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

$$
y \triangleq \alpha_2 v_* + \sum_{j=1}^{d_2-1} \beta_2^{(j)} v_j.
$$
Substitute $x$ and $y$ in (7) and (8) by their expansion, we then have

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

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, $xy^\top = \alpha_1\alpha_2 u_\ast v_\ast = M$, and $F(\alpha_1u_\ast, \alpha_2v_\ast) = 0$. Then we have global optima $(\alpha u_\ast, \frac{1}{\alpha}v_\ast)$ 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.

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

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

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

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

\[
(a^\top, b^\top) \nabla^2 F(x, y) \begin{pmatrix}
a \\
b
\end{pmatrix} = \sum_{i=1}^{d_2-1} (\beta_2^{(i)})^2 \|a\|^2 - 2a^\top Mb.
\]

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

\[
\square
\]

2.2 Proof of Lemma 2

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

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

One can verify that $\nabla^2 F(\alpha u_\ast, \frac{1}{\alpha}v_\ast)$ has eigenvalues $\alpha^2 + \frac{1}{\alpha^2}, \frac{1}{\alpha^2}, \frac{1}{\alpha^2}$. 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.

\[
\kappa \left( \nabla^2 F(\alpha u_\ast, \frac{1}{\alpha}v_\ast) \right) = \max\{\alpha^4, \frac{1}{\alpha^4}\} + 1.
\]
2.3 Proof of Lemma 3

Proof. Recall that \( M = u_*v_*^T \) and

\[
\begin{align*}
  x &\triangleq \alpha_1 u_* + \sum_{i=1}^{d_1-1} \sigma_1^{(i)} \bar{u}_i, \\
y &\triangleq \alpha_2 v_* + \sum_{j=1}^{d_2-1} \sigma_2^{(j)} \bar{v}_j.
\end{align*}
\]

By setting the gradient of \( \tilde{F} \) to zero we get

\[
\begin{align*}
  (\|x\|_2^2 + d_1 \sigma_1^2) y &= M^T x, \\
  (\|y\|_2^2 + d_2 \sigma_2^2) x &= My.
\end{align*}
\]

From the equations above, we can verify that \((x, y) = (0, 0)\) is a stationary point. Furthermore, left multiplying the equations above by \( \bar{v}_j \) and \( \bar{u}_i^T \) 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 \( v_*^T \) and \( u_*^T \) respectively, we will get

\[
\begin{align*}
  (\alpha_1^2 + d_1 \sigma_1^2) \alpha_1^2 &= \alpha_2 \alpha_1, \\
  (\alpha_2^2 + d_2 \sigma_2^2) \alpha_2^2 &= \alpha_1 \alpha_2.
\end{align*}
\]

Then, with some algebraic manipulations, we get

\[
\begin{align*}
  \alpha_1^2 &= \sqrt{\frac{d_1 \sigma_1^2}{d_2 \sigma_2^2} - d_1 \sigma_1^2}, \\
  \alpha_2^2 &= \sqrt{\frac{d_2 \sigma_2^2}{d_1 \sigma_1^2} - d_2 \sigma_2^2}.
\end{align*}
\]

Specifically, when \( d_1 \sigma_1^2 = \gamma^2 d_2 \sigma_2^2 \leq \gamma \), 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.

\[
\nabla^2 \tilde{F}(x, y) = \begin{pmatrix} \frac{\|y\|_2^2 + d_2 \sigma_2^2}{2y_\|y\|_2^2 - M^T} & \frac{2xy^T - M}{2y_\|y\|_2^2 - M^T} \end{pmatrix}.
\]

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

\[
\begin{align*}
  (a^T, b^T) \nabla^2 \tilde{F}(x_*, y_*) (\begin{array}{c} a \\ b \end{array}) &= (\alpha_2^2 + d_2 \sigma_2^2) ||a||_2^2 + (\alpha_1^2 + d_1 \sigma_1^2) ||b||_2^2 + 2(2\alpha_1 \alpha_2 - 1)(a^T u_*) (b^T v_*) \\
  &\geq (\alpha_2^2 + d_2 \sigma_2^2) ||a||_2^2 + (\alpha_1^2 + d_1 \sigma_1^2) ||b||_2^2 - 2[1 - \sqrt{d_1 \sigma_1^2 d_2 \sigma_2^2}] - 1 ||a||_2 ||b||_2 \\
  &> \sqrt{\frac{d_2 \sigma_2^2}{d_1 \sigma_1^2}} ||a||_2^2 + \sqrt{\frac{d_1 \sigma_1^2}{d_2 \sigma_2^2}} ||b||_2^2 - 2 ||a||_2 ||b||_2,
\end{align*}
\]

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

\[
\frac{d_2 \sigma_2^2}{d_1 \sigma_1^2} \|a\|_2^2 + \frac{d_1 \sigma_1^2}{d_2 \sigma_2^2} \|b\|_2^2 - 2 \|a\|_2 \|b\|_2 = \left( \frac{d_2 \sigma_2^2}{d_1 \sigma_1^2} \right)^2 \|a\|_2^2 - \left( \frac{d_1 \sigma_1^2}{d_2 \sigma_2^2} \right)^2 \|b\|_2^2.
\]

Thus, as long as \(d_1 \sigma_1^2 \neq d_2 \sigma_2^2 < 1\), we have \(\nabla^2 F_{reg}(x_*, y_*)\) 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

\[
(a^T, b^T) \nabla^2 F_{reg}(0, 0) \left( \begin{array}{c} a \\ b \end{array} \right) = d_2 \sigma_2^2 \|a\|_2^2 + d_1 \sigma_1^2 \|b\|_2^2 - 2(a^T u_*) (b^T v_*).
\]

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

\[\square\]

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 \(\sigma^2 = \mathbb{E}[\|\xi_1\|_2^2] = \mathbb{E}[\|\xi_2\|_2^2] = \|x_0\|_2^2 + \|y_0\|_2^2 \leq 1/\sigma^2\). For any \(\delta \in (0, 1)\), we take

\[
\eta \leq \eta_1 = C_1 \delta^2 \left(\log((d_1 + d_2)/\delta) \log(1/\delta)\right)^{-1},
\]

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

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

\[
\mathcal{A}_t = \left\{ \|\xi_{1, i}\|_2 \leq \sigma, \|\xi_{2, i}\|_2 \leq \sigma \left( \sqrt{2} \log((d_1 + d_2)/\eta \sigma^2) \right) \right\}, \forall \tau \leq t, i = 1, \ldots, d_1, j = 1, \ldots, d_2.
\]

By the concentration result of the maximum of Gaussian distribution, we have \(\mathbb{P}(A_1/\sigma^2) \geq 1 - \delta\). Moreover, we use \(\mathcal{H}_t\) to denote the event where the first \(t\) iterates \((x_{\tau}, y_{\tau})\) is bounded, i.e., \(\mathcal{H}_t = \left\{ \|x_{\tau}\|_2^2 + \|y_{\tau}\|_2^2 \leq \frac{2}{\sigma^2}, \forall \tau \leq t \right\}\).

Let \(\mathcal{F}_t = \mathcal{A}_t \cap \mathcal{H}_t\) 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 \(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2\).

\[
\mathbb{E}\left[\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2 \right| \mathcal{H}_t, \mathcal{A}_t, \mathcal{F}_t] \leq \left\{ \begin{array}{ll}
(1 - 2\eta\sigma^2) (\|x_t\|_2^2 + \|y_t\|_2^2) & - \eta (\|x_t\|_2^2 \|y_t\|_2^2 - \|x_t^T M y_t\|_2) + \eta^2 \mathbb{E}[\|\xi_1\|_2^2 \|\xi_2\|_2^2 (x_t + \xi_1, y_t + \xi_2, t) \|_2^2 + \|\nabla y F(x_t + \xi_1, y_t + \xi_2, t) \|_2^2] \mathbb{I}_{\mathcal{H}_t \cap \mathcal{A}_t} \\
(1 - 2\eta\sigma^2) (\|x_t\|_2^2 + \|y_t\|_2^2) & - \eta (\|x_t\|_2^2 \|y_t\|_2^2 - \|x_t^T y_t\|_2) + \eta^2 \mathbb{E}[\|\xi_1\|_2^2 \|\xi_2\|_2^2 (x_t + \xi_1, y_t + \xi_2, t) \|_2^2 + \|\nabla y F(x_t + \xi_1, y_t + \xi_2, t) \|_2^2] \mathbb{I}_{\mathcal{H}_t \cap \mathcal{A}_t} \\
(1 - 2\eta\sigma^2) (\|x_t\|_2^2 + \|y_t\|_2^2) & - \eta (\|x_t\|_2^2 \|y_t\|_2^2 - \|x_t^T y_t\|_2) + \eta^2 \mathbb{E}[\|\xi_1\|_2^2 \|\xi_2\|_2^2 (x_t + \xi_1, y_t + \xi_2, t) \|_2^2 + \|\nabla y F(x_t + \xi_1, y_t + \xi_2, t) \|_2^2] \mathbb{I}_{\mathcal{H}_t \cap \mathcal{A}_t} \\
(1 - 2\eta\sigma^2) (\|x_t\|_2^2 + \|y_t\|_2^2) & + \eta^2 C_2 \mathbb{I}_{\mathcal{H}_t \cap \mathcal{A}_t}
\end{array} \right.
\]

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

\[
\mathbb{E}\left[\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2 - 1/\sigma^2 \right| \mathcal{H}_t, \mathcal{A}_t, \mathcal{F}_t] \leq \left(1 - 2\eta \sigma^2\right) (\|x_t\|_2^2 + \|y_t\|_2^2 - 1/\sigma^2) \mathbb{I}_{\mathcal{H}_t \cap \mathcal{A}_t} \\
\leq \left(1 - 2\eta \sigma^2\right) (\|x_t\|_2^2 + \|y_t\|_2^2 - 1/\sigma^2) \mathbb{I}_{\mathcal{H}_{t-1} \cap \mathcal{A}_{t-1}}.
\]

\[\square\]
Then when with probability at least
\[ \Pr(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2 - 1/\sigma^2) \leq \delta, \] 
\( G_t \mathbf{1}_{\mathcal{H}_{t-1} \cap A_{t-1}} \) 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 \( G_t \mathbf{1}_{\mathcal{H}_{t} \cap A_{t}} \) and \( \mathbb{E}[G_t \mathbf{1}_{\mathcal{H}_{t} \cap A_{t}}, \mathcal{F}_t] \).

\[
d_{t+1} = |G_{t+1} \mathbf{1}_{\mathcal{H}_{t} \cap A_{t}} - \mathbb{E}[G_{t+1} \mathbf{1}_{\mathcal{H}_{t} \cap A_{t}} | \mathcal{F}_t]|^2 \leq (1 - 2\eta \sigma^2)^{t-1} \left( \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}},
\]

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

\[
\Pr(G_t \mathbf{1}_{\mathcal{H}_{t-1} \cap A_{t-1}} - G_0 \geq \mathbb{E}(1)r_t \left( \log \frac{1}{r_t^2 \delta} \right) \leq O(\eta^2 \delta).
\]

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

\[
(\|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2) \mathbf{1}_{\mathcal{H}_{t} \cap A_{t}} \leq 1/\sigma^2 + (1 - \eta \sigma^2)^t (\|x_0\|_2^2 + \|y_0\|_2^2 - 1/\sigma^2)
\]

\[
+ O(1)(1 - \eta \sigma^2)^t r_t (\log \frac{1}{r_t^2 \delta})^{\frac{1}{2}}
\]

\[
\leq 1/\sigma^2 + O \left( \sqrt{\eta \sigma^2} \left( \log \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) \left( \log \frac{1}{\eta^2 \delta} \right)^{\frac{1}{2}}
\]

\[
\leq 2/\sigma^2,
\]

when \( \eta = O \left( (\log \frac{d_1 + d_2}{\delta} \log \frac{1}{\delta})^{-1} \right) \). In order to satisfy \( \eta \leq 1/C_2 \) at the same time, we take \( \eta \leq \eta_1 = O \left( \sigma^6 (1 - \eta \sigma^2)^{-1} \right) \) to make sure that all inequalities above hold.

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

\[
\Pr(\mathcal{A}_{1/\eta^2}) \geq 1 - \delta.
\]

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

\[ \square \]

### 3.2 Proof of Lemma 4

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

\[
E_m \left[ \frac{\langle \zeta_{\ell,1}^{(t)} \rangle^2}{\sigma^2} \right] = \frac{1}{\sigma^2} \left( \frac{d_1 + d_2}{\eta} \right)
\]

where \( d_1 + d_2 \) is the number of elements in the vector \( \zeta_{\ell}^{(t)} \), and \( \eta \) is the step size. By taking the expectation, we have

\[
E_m \left[ \frac{\langle \zeta_{\ell,1}^{(t)} \rangle^2}{\sigma^2} \right] \leq \frac{d}{\eta^2} \leq \frac{d}{\delta^2}
\]

and the proof of Lemma 4 is complete.
We next show that with probability at least
\[
\Pr(\|\beta_{t+1}\|_2^2 \geq \frac{\eta C_2}{\sigma^2}) \leq \frac{\eta C_2}{\sigma^2}.
\]
Denote \(H_t = \left\{ \forall \tau \leq t, \|\beta_{\tau}\|_2^2 \geq \frac{\eta C_2}{\sigma^2} \right\}\). By (11), we have
\[
\mathbb{E}[G_{t+1}\mathbb{1}_{H_t}\mathbb{1}_{E_t}] \leq G_t\mathbb{1}_{H_t} \leq G_t\mathbb{1}_{E_t}.
\]
Hence, by Markov inequality we have
\[
P(E_t) = \Pr\left(\|\beta_{\tau}\|_2^2 \mathbb{1}_{E_t} \geq \frac{\eta C_2}{\sigma^2}\right) \leq \frac{\mathbb{E}[\|\beta_{\tau}\|_2^2 \mathbb{1}_{E_t}]}{\frac{\eta C_2}{\sigma^2}} \leq \left(1 - 2\eta \sigma^2\right) t \left(\|\beta_{\tau}\|_2^2 - \frac{\eta C_2}{2\sigma^2}\right) + \frac{\eta C_2}{2\sigma^2} \leq \left(1 - 2\eta \sigma^2\right) t \frac{\eta C_2}{2\sigma^2} + \frac{1}{2} \leq \frac{3}{4},
\]
when \(t \geq \frac{\log(4/3)}{\eta^2} \log \frac{8}{C_2^2 \eta}\). We take \(t = \frac{1}{\eta^2} \log \frac{8}{C_2^2 \eta}\) to make sure the inequality above holds. Thus with probability at least \(\frac{1}{4}\), there exists a \(\tau \leq \frac{1}{\eta^2} \log \frac{8}{C_2^2 \eta}\), such that \(\|\beta_{\tau}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2}\). Thus, with probability at least \(1 - \delta\), we can find a \(\tau\) such that \(\|\beta_{\tau}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2}\), where
\[
\tau \leq \tau_1 = \frac{1}{\log(4/3)\eta^2} \log \frac{8}{C_2^2 \eta} \log \frac{1}{\delta} = \frac{1}{\log(4/3)\eta^2} \left(\log \frac{8}{C_2^2} + \log \frac{1}{\delta}\right) \log \frac{1}{\delta} = O\left(\frac{1}{\eta^2} \log \frac{1}{\eta} \log \frac{1}{\delta}\right).
\]
We next show that with probability at least \(1 - \delta\), for \(\forall t \geq \tau_1\), we have \(\|\beta_{t+1}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2}\). This can be done following the similar lines to the proof of Lemma 1. We first restart the counter of the time and assume \(\|\beta_{t,0}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2}\). Denote \(H_t = \left\{ \forall \tau \leq t, \|\beta_{\tau}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2} \right\}\). Then by (11), we have
\[
\mathbb{E}[G_{t+1}\mathbb{1}_{H_{t+1},\mathbb{1}_{H_t}}\mathbb{1}_{E_t}] \leq G_t\mathbb{1}_{H_{t+1} \cap H_t} \leq G_t\mathbb{1}_{H_{t+1} \cap H_t}.
\]
The difference of \(G_{t+1}\mathbb{1}_{H_{t+1} \cap H_t} = \mathbb{E}[G_{t+1}\mathbb{1}_{H_{t+1} \cap H_t} | \mathcal{F}_t] - \mathbb{E}[G_{t+1}\mathbb{1}_{H_{t+1} \cap H_t} | \mathcal{F}_t]\) can be easily bounded as follows
\[
D_{t+1} = \left| G_{t+1}\mathbb{1}_{H_{t+1} \cap H_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{H_{t+1} \cap H_t} | \mathcal{F}_t] \right| = C_2 \eta^2 \sigma^{-3} \left(\log \frac{d_1 + d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log \frac{1}{\delta}\right)^{\frac{1}{2}},
\]
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
\[
\|\beta_{t,1}\|_2^2 \leq \frac{2\eta C_2}{\sigma^2},
\]
for any \(\tau_1 \leq t, 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 \(|\beta_{1,t}|^2 \leq 2\frac{\eta C_2}{\sigma^2}\) and \(|\beta_{2,t}|^2 \leq 2\frac{\eta C_2}{\sigma^2}\) 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 \(\tau \leq \tau_2\), such that
\[
x_t^\top M y_t \geq 9 \frac{\eta C_2}{\sigma^4},
\]
where \(\tau_2 = \frac{5}{\eta^2} \log \frac{4\eta^2}{\sigma^2} \log \frac{1}{\delta}\).

Proof. Let’s define the event \(\mathcal{H}_t = \{|x_t|^2 M y_t \leq 9 \frac{\eta C_2}{\sigma^2}, \forall \tau \leq t\}\). Following the proof of Lemma 1, we can refine the bound on the conditional expectation of \(||x_{t+1}|^2 + |y_{t+1}|^2|\) given \(\mathcal{H}_t\). Specifically, we have
\[
\mathbb{E}\left[\left(|x_{t+1}|^2 + |y_{t+1}|^2\right)\mathbb{I}_{\mathcal{H}_t}|\mathcal{F}_t\right] = \left\{(1 - 2\eta \sigma^2) \left(|x_t|^2 + |y_t|^2\right) - 4\eta \left(|x_t|^2 |y_t|^2 - x_t^\top M y_t\right)\right\} \mathbb{I}_{\mathcal{H}_t},
\]
\[
+ \eta^2 \mathbb{E}_{\xi_i, \xi_j, \xi_o} \left|\nabla_x F(x_t + \xi_1, y_t + \xi_2, t)\right|_2 + \left|\nabla_y F(x_t + \xi_1, y_t + \xi_2, t)\right|_2 \mathbb{I}_{\mathcal{H}_t},
\]
\[
\leq \left\{(1 - 2\eta \sigma^2) \left(|x_t|^2 + |y_t|^2\right) + 34C_2\eta^2 \sigma^{-4}\right\} \mathbb{I}_{\mathcal{H}_t}.
\]
Denote \(G_t = (1 - 2\eta \sigma^2)^{-t} \left(|x_t|^2 + |y_t|^2 - 17C_2\eta \sigma^{-6}\right)\). Then we have
\[
\mathbb{E}[G_{t+1} \mathbb{I}_{\mathcal{H}_{t+1}}] \leq G_t \mathbb{I}_{\mathcal{H}_t}, \leq G_t \mathbb{I}_{\mathcal{H}_{t+1}}.
\]
Thus, by Markov inequality, we have
\[
\mathbb{P}(\left(|x_t|^2 + |y_t|^2\right) \mathbb{I}_{\mathcal{H}_{t+1}} \geq 34C_2\eta \sigma^{-6}) \leq \frac{(1 - 2\eta \sigma^2)^t \left(|x_0|^2 + |y_0|^2 - 17C_2\eta \sigma^{-6}\right) + 17C_2\eta \sigma^{-6}}{34C_2\eta \sigma^{-6}} \leq \frac{3}{4},
\]
when \(t \geq \frac{1}{3\eta \sigma^2} \log \frac{4\eta^2}{\sigma^2} \log \frac{1}{\delta}\). Thus with probability at least \(1 - \delta\), there exists a \(\tau \leq \tau_2 = \frac{5}{\eta^2} \log \frac{4\eta^2}{\sigma^2} \log \frac{1}{\delta}\), such that
\[
(|x_{\tau}|^2 + |y_{\tau}|^2) \mathbb{I}_{\mathcal{H}_{\tau+1}} \leq 34C_2\eta \sigma^{-6}.
\]
Following the exactly same proof of Lemma 1, we can show that with probability at least \(1 - \delta\), for all \(\tau_2 \leq t \leq T_1 = O\left(\frac{1}{\eta^2}\right)\), we have
\[
\left(|x_t|^2 + |y_t|^2\right) \mathbb{I}_{\mathcal{H}_{t+1}} \leq (17 + 34)C_2\eta \sigma^{-6} + O\left(\sqrt{\eta \sigma^{-2}} \left(\left(\log \frac{d_1 + d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log \frac{1}{\delta}\right)^{\frac{1}{2}}\right) \right) \left(\log \frac{1}{\eta^2 \delta}\right)^{\frac{1}{2}}
\]
\[
= C_4' \eta \sigma^{-12} + C_5' \sqrt{\eta \sigma^{-2}} \left(\left(\log \frac{d_1 + d_2}{\eta}\right)^{\frac{1}{2}} + \left(\log \frac{1}{\delta}\right)^{\frac{1}{2}}\right) \left(\log \frac{1}{\eta^2 \delta}\right)^{\frac{1}{2}}
\]
\[
\leq C_6' \left(\log \frac{d_1 + d_2}{\delta} \log \frac{1}{\delta}\right)^{-1} + C_7' \sigma^4,
\]
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 \(t \leq \tilde{\tau}_2 = O\left(\frac{1}{\eta^2} \log \frac{1}{\eta} \log \frac{1}{\delta}\right)\), by the analysis above we have \(|x_t|^2 + |y_t|^2 \leq C_6' \left(\log \frac{d_1 + d_2}{\delta} \log \frac{1}{\delta}\right)^{-1} + C_7' \sigma^4\) holds for \(\forall t \leq \tilde{\tau}_2\).
We next give a lower bound on the conditional expectation of

\[ |\alpha_{1,t+1} + \alpha_{2,t+1}| - |\alpha_{1,t} + \alpha_{2,t}| \]

Thus, we know that, with probability at least

\[ C'_8 \eta(\sigma_1 + \sigma_2), \]

where \( C'_8 \) is some positive constant. This means we can find a \( \tau = O(\log \frac{1}{\delta}) \leq \tilde{\tau}_2 \), such that \( |\alpha_{1,\tau} + \alpha_{2,\tau}| > C'_8 \eta(\sigma_1 + \sigma_2) \), 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 \( |\alpha_{1,t+1} + \alpha_{2,t+1}| \).

\[
E \left[ |\alpha_{1,t+1} + \alpha_{2,t+1}| \right] \geq E \left[ (|\alpha_{1,t+1} + \alpha_{2,t+1}|) \right] = \left(1 + q(1 - \sigma^2 - \alpha_{1,t}\alpha_{2,t}) \right) (|\alpha_{1,t} + \alpha_{2,t}| - \eta (\|\beta_{1,t}\| \|\alpha_{2,t} + \beta_{2,t}\|^2)) \]

\[
\geq \left(1 + \eta(1 - \sigma^2 - \alpha_{1,t}\alpha_{2,t}) \right) |\alpha_{1,t} + \alpha_{2,t}| - 2C_9 \frac{\eta^{1,2}C_2}{\sigma^3} \left( \left( \log \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right)
\]

where \( C_9 \) is some positive constant. This is equivalent to the following inequality:

\[
E \left[ |\alpha_{1,t+1} + \alpha_{2,t+1}| - 4C_9 \frac{\eta^{1,2}C_2}{\sigma^3} \left( \left( \log \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) \right] \geq \left(1 + \frac{1}{2} \eta \right) \left( |\alpha_{1,t} + \alpha_{2,t}| - 4C_9 \frac{\eta^{1,2}C_2}{\sigma^3} \left( \left( \log \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) \right)
\]

\[
\geq \left(1 + \frac{1}{2} \eta \right)^{t+1} \left( |\alpha_{1,0} + \alpha_{2,0}| - 4C_9 \frac{\eta^{1,2}C_2}{\sigma^3} \left( \left( \log \frac{d_1 + d_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) \right)
\]

\[
\geq \left(1 + \frac{1}{2} \eta \right)^{t+1} \left( C_9 \eta(d_1 + d_2) - 4C_9 \frac{\eta^{1,2}C_2}{\sigma^3} \left( \left( \log \frac{\sigma_1 + \sigma_2}{\eta} \right)^{\frac{1}{2}} + \left( \log \frac{1}{\delta} \right)^{\frac{1}{2}} \right) \right)
\]

\[
\geq C_{10}' \eta(\sigma_1 + \sigma_2) \exp t\eta/2
\]

\[
\geq C_{10}' \eta(\sigma/\sqrt{d_1} + \sigma/\sqrt{d_2}) \frac{1}{\eta\sigma}
\]

\[
= C_{10}'(1/\sqrt{d_1} + 1/\sqrt{d_2}),
\]

where \( C_{10}' \) is some positive constant and last inequality holds when \( t \geq 2 \log \frac{1}{\eta \delta} \). Note that here we still take \( \eta = O \left( \sigma^{12} \left( \log \frac{d_1 + d_2}{\eta} \log \frac{1}{\delta} \right)^{-1} \right) \), 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 \( \delta \) and \( \sigma \) go to zero (which guarantees that \( \eta \) 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, \( \|x_{t+1}\|^2 + \|y_{t+1}\|^2 \) stays in a very small neighborhood around zero, which makes \( \alpha_{1,t+1}^2 + \alpha_{2,t+1}^2 \leq \|x_{t+1}\|^2 + \|y_{t+1}\|^2 \) also very small. This implies that \( |\alpha_{1,t+1} + \alpha_{2,t+1}| \) can be arbitrarily small as long as we make \( \delta \) and \( \sigma \) small enough, which is the desired contradiction.

Thus, we know that, with probability at least \( 1 - \delta \), there exists a \( t \) satisfying \( \tau_2 \leq t \leq \tau_2 + \tilde{\tau}_2 \leq 2\tau_2 \), such that \( \mathcal{H}_t \) does not hold. Re-scale \( \tau_2 \) 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 $x_t^\top My_0 \geq \frac{C_2}{\eta^2}$. We take $\eta$ as in Lemma 5. Then there almost surely exists a $\tau \leq \tau'_2 = \frac{2}{\eta} \log \frac{2a^2}{C_2^2}$, such that

$$x_t^\top My_\tau \geq \frac{1}{2} + \sigma^2.$$ 

**Proof.** Suppose $x_t^\top My_t \leq \frac{1}{2} + \sigma^2$ holds for all $t \leq \tau'_2$. We next give a lower bound on the conditional expectation of $|\alpha_{1,t+1} + \alpha_{2,t+1}|$.

$$\mathbb{E} \left[ |\alpha_{1,t+1} + \alpha_{2,t+1}| \big| F_t \right] \geq \mathbb{E} \left[ (\alpha_{1,t+1} + \alpha_{2,t+1}) | F_t \right] = \left( 1 + \eta(1 - \sigma^2 - \alpha_{1,t}\alpha_{2,t}) \right) (\alpha_{1,t} + \alpha_{2,t}) - \eta \left( \|\beta_{1,t}\|_2^2 \alpha_{2,t} + \|\beta_{2,t}\|_2^2 \alpha_{1,t} \right) \geq \left( 1 + \frac{1}{2} \eta \right) |\alpha_{1,t} + \alpha_{2,t}| - \frac{4\eta^2 C_2}{\sigma^4}.$$ 

This implies that

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

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

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

when $t = \tau'_2 = \frac{2}{\eta} \log \frac{2a^2}{C_2^2}$. On the other hand,

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

which leads to a contradiction unless

$$\mathbb{P}(x_t^\top My_t \leq \frac{1}{2} + \sigma^2, \forall t \leq \tau'_2) = 0.$$ 

We prove the result. \(\square\)

**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 $\tau$ such that for some positive constant $c < \frac{1}{2}$

$$x_\tau^\top My_\tau > 2c,$$

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

$$(x_t, y_t) \in \{(x, y) | x^\top My \geq c\}.$$

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 $1 - \delta$,

$$(\alpha_{1,t}, \alpha_{2,t}) \in \{(\alpha_1, \alpha_2) | (\alpha_1 \alpha_2 - 1)^2 < (1 - c)^2\}.$$
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 \(T_1 = O(\frac{1}{c^2})\) steps,
\[
\|\xi_{k,t}\|_\infty \leq \bar{\sigma} \triangleq \sigma \left( \sqrt{2\log((d_1 + d_2)\eta^{-2})} + \sqrt{\log(1/\delta)} \right).
\]

**Case (i):** By plugging (5) and (6) to (2), we get
\[
\alpha_{1,t+1}\alpha_{2,t+1} = \alpha_{1,t}\alpha_{2,t} - \eta \left( \alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2_2) - \alpha_{2,t} + g_x) + \alpha_{1,t}(\alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2_2) - \alpha_{1,t} + g_y) \right) \\
+ \eta^2(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2_2) - \alpha_{2,t} + g_x)(\alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2_2) - \alpha_{1,t} + g_y) \\
= 2c + (\alpha_{1,t}\alpha_{2,t} - 2c) - \eta[\alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2_2) - \alpha_{2,t} + g_x) + \alpha_{1,t}(\alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2_2) - \alpha_{1,t} + g_y)] \\
+ \eta(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2_2) - \alpha_{2,t} + g_x)(\alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2_2) - \alpha_{1,t} + g_y),
\]

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 \(\eta\) 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 \(O\left(\frac{1}{\sigma} \left( \sqrt{\log(d_1 + d_2)} + \sqrt{\frac{1}{\delta}} \right) \right)\) from the following upper bound:
\[
|g_x| \leq \bar{\sigma}(\alpha_{2,t}^2 + 2\alpha_{2,t}\alpha_{1,t} + 1 + \frac{2\eta C_2}{\sigma^2}) + \bar{\sigma}^2(2\alpha_{2,t} + \alpha_{1,t}) + \bar{\sigma}^3, \\
|g_y| \leq \bar{\sigma}(\alpha_{1,t}^2 + 2\alpha_{1,t}\alpha_{2,t} + 1 + \frac{2\eta C_2}{\sigma^2}) + \bar{\sigma}^2(2\alpha_{1,t} + \alpha_{2,t}) + \bar{\sigma}^3.
\]

It is easy to show that for properly selected
\[
\eta \leq \bar{\eta}_1 = O \left( \left( \frac{1}{\sigma^2} + \sqrt{\log(d_1 + d_2)} + \sqrt{\frac{1}{\delta}} \right)^{-1} \right),
\]
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 \(2c < \alpha_{1,t}\alpha_{2,t} < 2 - 2c\).

\[E_x[A_t] = E_x[\alpha_{2,t} < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_t > + \alpha_{1,t} < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_t >].\]

By plugging (5) and (6) in the above equation, we have
\[
\bar{E}_x[\alpha_{2,t}] = \alpha_{2,t}(\alpha_{1,t}(\alpha_{2,t}^2 + \|\beta_{2,t}\|^2_2) - \alpha_{2,t} + \alpha_{1,t}\sigma^2) + \alpha_{1,t}(\alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2_2) - \alpha_{1,t} + \alpha_{2,t}\sigma^2) \\
= (\alpha_{2,t}(\alpha_{1,t} - 1)(\alpha_{2,t}^2 + \alpha_{1,t}^2) + \alpha_{2,t}\alpha_{1,t}(2\sigma^2 + \|\beta_{1,t}\|^2_2 + \|\beta_{2,t}\|^2_2)).
\]

Again, by plugging (5) and (6) to (3) and taking expectation conditioning on \(F_t\), when \(\alpha_{1,t}\alpha_{2,t} > c\), we will get
\[
\bar{E}(\alpha_{1,t+1}\alpha_{2,t+1} - 1)^2[F_t] = \left((\alpha_{1,t}\alpha_{2,t} - 1)^2 - 2\eta E_x[A_t]\right)(\alpha_{1,t}\alpha_{2,t} - 1) \\
+ \eta^2E_x[A_t^2] + \eta^4 E_x[B_t^2] - 2\eta^3 E_x[A_t B_t] + 2\eta^2 E_x[B_t](\alpha_{1,t}\alpha_{2,t} - 1) \\
= (1 - 2\eta(\alpha_{2,t} + \alpha_{1,t}^2))(\alpha_{1,t}\alpha_{2,t} - 1)^2 \\
- 2\eta(\alpha_{1,t}(\alpha_{1,t}\alpha_{2,t} - 1)(2\sigma^2 + \|\beta_{1,t}\|^2_2 + \|\beta_{2,t}\|^2_2) \\
+ \eta^2 E_x[A_t^2] + \eta^4 E_x[B_t^2] - 2\eta^3 E_x[A_t B_t] + 2\eta^2 E_x[B_t](\alpha_{1,t}\alpha_{2,t} - 1) \\
\leq (1 - 4\eta c)(\alpha_{1,t}\alpha_{2,t} - 1)^2 + 2\eta^2(2\sigma^2 + \frac{2\eta C_2}{\sigma^2}) + \bar{C}_1 \eta^2 \\
\leq (1 - 4\eta c)(\alpha_{1,t}\alpha_{2,t} - 1)^2 + \eta(\sigma^2 + \frac{2\eta C_2}{\sigma^2}) + \bar{C}_1 \eta^2,
\]
where \(\bar{C}_1 = O \left( \frac{1}{\sigma^2} \left( \sqrt{\log(d_1 + d_2)} \right) \right),\) if \(\alpha_{1,t}\alpha_{2,t}\) is of constant order. Note that here we choose \(\eta \leq \bar{\eta}_1\) as mentioned in (i). Denote \(\gamma \triangleq \frac{\eta(\sigma^2 + \frac{2\eta C_2}{\sigma^2}) + \bar{C}_1 \eta^2}{4\eta c}\), the inequality above can be re-expressed as
\[
E_x[(\alpha_{1,t+1}\alpha_{2,t+1} - 1)^2 - \gamma] [F_t] \leq (1 - 4\eta c) (\alpha_{1,t}\alpha_{2,t} - 1)^2 - \gamma.
\]
\[\text{ (12)}\]
We denote \( G_t \equiv (1 - 4\eta c)^{-t}((\alpha_{1,t}\alpha_{2,t} - 1)^2 - \gamma) \) and \( E_t \equiv \{ \forall \tau \leq t : (\alpha_{1,\tau}\alpha_{2,\tau} - 1)^2 < (1 - c)^2 \} \). 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

\[
E[G_{t+1}\mathbb{1}_{E_t}\mathbb{F}_{\tau}] \leq G_t\mathbb{1}_{E_t} \leq G_t\mathbb{1}_{E_{t-1}}.
\]

This means \( \{ G_t\mathbb{1}_{E_{t-1}} \} \) is a supermartingale. To use Azuma’s inequality, we need to bound the following difference

\[
d_{t+1} \equiv |G_{t+1}\mathbb{1}_{E_t} - E[G_{t+1}\mathbb{1}_{E_t}\mathbb{F}_{\tau}]|
\]

\[
= (1 - 4\eta c)^{-t-1}((\alpha_{1,t+1}\alpha_{2,t+1} - 1)^2 - E[(\alpha_{1,t+1}\alpha_{2,t+1} - 1)^2]\mathbb{F}_{\tau}]|\mathbb{1}_{E_t}
\]

\[
\leq (1 - 4\eta c)^{-t-1}\bar{C}_2,
\]

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

\[
\mathbb{P}\left(G_t\mathbb{1}_{E_{t-1}} - G_0 \geq O(1)r_t\log^2\left(\frac{1}{\eta^2\delta}\right)\right) \leq \exp\left(-\frac{O(1)r_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^{t} d_i^2}\right) = O\left(\eta^2\delta\right).
\]

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

\[
((\alpha_{1,t}\alpha_{2,t} - 1)^2 - \gamma)\mathbb{1}_{E_{t-1}}, < (1 - 4\eta c)^t((\alpha_{1,0}\alpha_{2,0} - 1)^2 + O(1)r_t\log^2\left(\frac{1}{\eta^2\delta}\right))
\]

\[
< (\alpha_{1,0}\alpha_{2,0} - 1)^2 + O\left(1 - 4\eta c\right)r_t\log^2\left(\frac{1}{\eta^2\delta}\right)
\]

\[
< (1 - c)^2 + (1 - 2c)^2 - (1 - c)^2
\]

\[
+ O\left(\frac{\sqrt{\eta}}{\sigma^2} \left(\sqrt{\log \frac{d_1 + d_2}{\eta^2}} + \sqrt{\frac{1}{\delta}}\right)\log^2\left(\frac{1}{\eta^2\delta}\right)\right) + \gamma < (1 - c)^2,
\]

as \( (1 - 2c)^2 - (1 - c)^2 \) is some negative constant, by choosing \( \eta \leq \bar{\eta}_2 = O\left(\sigma^4(\log \frac{1}{\delta})^{-1}\left(\log \left(\frac{d_1 + d_2}{\delta}\right)\right)^{-1}\right) \) and \( \sigma \) small enough, we can make sure the sum of last four terms is negative. Now we know that if \( E_{t-1} \) holds, \( E_t \) 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 \( O\left(\frac{1}{\eta^2}\right) \) steps w.p. at least \( 1 - \delta \).

\[\square\]

### 3.4 Proof of Lemma 6

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

**Lemma 5.** Suppose \( \forall t \leq T_1 \equiv O\left(\frac{1}{\eta^2}\right), \)

\[
(x_t, y_t) \in \{(x, y) : x^\top M y > c\},
\]

then \( \forall \delta \in (0, 1), \) with probability at least \( 1 - \delta \), there exists a time step \( \tau \leq \tau_3 \equiv O\left(\frac{1}{\eta^2} \log \frac{1}{\sigma} \log \frac{1}{\delta}\right) \) such that

\[
(x_{\tau}^\top M y_{\tau} - 1)^2 < 4\gamma,
\]

where \( \gamma = O(\sigma^2) \).
Lemma 6. Suppose there exists a time step $\tau \leq \tau_3 \triangleq O\left(\frac{1}{\eta} \log \frac{1}{\eta^2} \log \frac{1}{3}\right)$ such that

$$(x_t^\top My - 1)^2 < 4\gamma,$$

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

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

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

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

Here follows proof of Lemmas 5 and 6.

**Proof.** Define $H_t \triangleq \{\forall \tau \leq t : (\alpha_{1,\tau} - 1)^2 \geq 4\gamma\}$, for $t > \frac{\log(\frac{\alpha_{1,\alpha_{2,0} - 1}^2)}{4\eta_c}}{-\eta_c}$, we have

$$4\gamma E[I_{H_t}] \leq E[(\alpha_{1,\tau} - 1)^2 - \gamma] \leq (1 - 4\eta c)^t ((\alpha_{1,0} - 1)^2 - \gamma) + \gamma < 2\gamma,$$

where the first inequality comes from the definition of $H_t$ and the second one comes from (12). Thus, if we choose $t = \max(\frac{\log(\frac{\alpha_{1,\alpha_{2,0} - 1}^2)}{4\eta_c}}{-\eta_c}, \log(\frac{1}{\gamma^2})), \log(\frac{1}{\gamma^2})) = \left(\frac{1}{\eta} \log \frac{\log(\frac{1}{\gamma^2})}{\gamma^2}\right)$, we have

$$\Pr(H_{\tau_3}) < \left(\frac{1}{2}\right)^{\log(\frac{1}{\gamma^2})} = \delta.$$

Thus, w.p. at least $1 - \delta$, there exists a $\tau \leq \tau_3$ s.t. $(\alpha_{1,\alpha_{2,0} - 1}^2) < 4\gamma$. Here, we finish the proof of Lemma 5.

Without loss of generality, we place the time origin at $\tau$, i.e. $(\alpha_{1,0} - 1)^2 < 4\gamma$. We next prove Lemma 6. Denote $A \triangleq \{(\alpha_{1,0} - 1)^2 \leq 6\gamma\}$ and $A_t \triangleq \{\forall \tau \leq t : (\alpha_{1,\tau} - 1)^2 \leq 6\gamma < (1 - \eta c)^t \leq 6\gamma$. Thus, without any assumption, we can get

$$E[G_{t+1}I_{A_t} | F_t] \leq G_t I_{A_t} \leq G_t I_{A_{t-1}}.$$

This means $\{G_t I_{A_{t-1}}\}$ is a supermartingale. To use Azuma’s inequality, we need to bound the following difference

$$d_{t+1} \triangleq |G_{t+1} I_{A_t} - E[G_{t+1} I_{A_t} | F_t]|$$

$$= (1 - 4\eta c)^t (\alpha_{1,\tau} - 1)^2 - E[\alpha_{1,\tau} - 1)^2 | F_t | I_{A_t}$$

$$\leq (1 - 4\eta c)^t \frac{C_3}{3},$$

where $C_3 = O(\sqrt{\gamma C_2})$. Further define $r_t \triangleq \sqrt{\sum_{i=1}^{t} d_i^2}$, by Azuma’s inequality we have

$$\Pr\left(G_t I_{A_{t-1}} - G_0 \geq O(1) r_t \log \frac{1}{\eta^2} \log \frac{1}{\eta^2}\right) \leq \exp\left(\frac{-O(1) r_t \log \frac{1}{\eta^2}}{2 \sum_{i=1}^{t} d_i^2}\right) = O(\eta^2 \delta).$$

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

$$(\alpha_{1,\tau} - 1)^2 \leq (1 - 4\eta c)^t ((\alpha_{1,0} - 1)^2 + O(1) r_t \log \frac{1}{\eta^2} \log \frac{1}{\eta^2})$$

$$< (\alpha_{1,0} - 1)^2 + O(1) r_t \log \frac{1}{\eta^2} \log \frac{1}{\eta^2}$$

$$< 4\gamma + O\left(\frac{\sqrt{\gamma}}{\eta^2} \log \frac{1}{\eta^2} \log \frac{1}{\eta^2} \log \frac{1}{\eta^2}\right).$$
When $E_{t-1}$ holds, w.p. at least $1 - O(\eta^2 \delta)$

$$(\alpha_{1,t} \alpha_{2,t} - 1)^2 < 5\gamma + O\left(\frac{C_3}{\sqrt{\eta}} \log^\frac{1}{\eta^2 \delta}\right) < 6\gamma,$$

by choosing $\eta \leq \tilde{\eta}_3 = O\left(\sigma^6 (\log \frac{1}{\Delta})^{-1} (\log \frac{d_1 + d_2}{\delta})^{-1}\right)$ we can guarantee the last term is smaller than $\gamma$. Now we know that if $A_{t-1}$ holds, $A_t$ 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 $O\left(\frac{1}{\sigma^2}\right)$ steps w.p. at least $1 - \delta$.

3.5 Proof of Lemma 7

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

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

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

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

for all $t$'s such that $\tau_4 \leq t \leq T_1$, where $T_1 = O\left(\frac{1}{\eta \sigma}\right)$ and $\tau_4 = O\left(\frac{1}{\eta \sigma} \log \frac{1}{\eta \sigma} \log \frac{1}{\Delta}\right)$.

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

**Lemma 8.** Suppose $\forall t \leq T_1 = O\left(\frac{1}{\eta \sigma}\right)$,

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

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

$$(x_t^\top u_*)^2 - (y_t^\top v_*)^2 < 4\Delta,$$

where $\Delta = O\left(\frac{\eta}{\sigma^2}\right)$.

**Proof.**

$$\mathbb{E}_x[D_{t+1}] = \mathbb{E}_x[\alpha_{1,t} < \nabla_x F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), u_*, > - \alpha_{2,t} < \nabla_y F(x_t + \xi_{1,t}, y_t + \xi_{2,t}), v_*, >]$$

$$= \alpha_{1,t}(\alpha_{1,t}^2 + \|\beta_{2,t}\|^2) - \alpha_{2,t} - \alpha_{1,t} \sigma^2 - \alpha_{2,t}(\alpha_{2,t}^2 + \|\beta_{1,t}\|^2) - \alpha_{1,t} + \alpha_{2,t} \sigma^2$$

$$= \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2) + (\alpha_{1,t}^2 - \alpha_{2,t}^2)(\beta_{1,t}^2 - \alpha_{2,t}^2)$$

Thus, we have

$$\mathbb{E}_x[D_{t+1}](\alpha_{1,t}^2 - \alpha_{2,t}^2) = \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 + (\alpha_{1,t}^2 - \alpha_{2,t}^2)(\alpha_{1,t}^2 - \alpha_{2,t}^2) - \alpha_{2,t}^2$$

$$\geq \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 2\alpha_{1,t}^2 \alpha_{2,t}^2 \frac{2\eta C_2}{\sigma^2}$$

$$> \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 2(1 + \sqrt{6\gamma})^2 \frac{2\eta C_2}{\sigma^2}$$

$$> \sigma^2(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 3 \frac{2\eta C_2}{\sigma^2}.$$

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

$$\mathbb{E}[(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 | \mathcal{F}_t] = (\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - 4\eta \mathbb{E}_x[D_{t+1}](\alpha_{1,t}^2 - \alpha_{2,t}^2)$$

$$+ 4\eta^2 \mathbb{E}_x[D_{t+1}^2] + \eta^4 \mathbb{E}_x[F_{t+1}^2] - 4\eta^3 \mathbb{E}_x[D_{t+1} F_t] + 2\eta^2 \mathbb{E}_x[F_{t+1}^2](\alpha_{1,t}^2 - \alpha_{2,t}^2)$$

$$\leq (1 - 4\eta \sigma^2)(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 + 24\eta \frac{\eta C_2}{\sigma^2} + C_4 \eta^2.$$
where $\bar{C}_4 = O(\frac{1}{\sigma^2} \left( \sqrt{\log \frac{d_1 + d_2}{\eta^2}} + \sqrt{\log \frac{1}{\delta}} \right)^2 )$. Denote $\Delta \triangleq \frac{24\epsilon_3 \bar{C}_4 + \bar{C}_4 \eta^2}{4\eta^2}$. The inequality above can be re-expressed as

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

Denote $\mathcal{B}_t \triangleq \{\forall \tau \leq t : (\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 \geq 4\Delta \}$, for $t > \frac{\log(\frac{\epsilon_3 C^2}{4\eta^2})}{4\eta^2}$, we have

$$4\Delta E[|\mathcal{B}_t|] \leq E[(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - \Delta] \leq (1 - 4\eta \sigma^2)(\alpha_{1,0}^2 - \alpha_{2,0}^2)^2 - \Delta + \Delta < 2\Delta,$$

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

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

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

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

$$((x_{\tau}^{\top} u_{\tau})^2 - (y_{\tau}^{\top} v_{\tau})^2)^2 < 4\Delta,$$

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

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

**Proof.** Without loss of generality, we place the time origin at $\tau$, i.e. $(\alpha_{1,0}^2 - \alpha_{2,0}^2)^2 < 4\Delta$. Denote $\mathcal{D} \triangleq \{(\alpha_{1,0}^2 - \alpha_{2,0}^2)^2 < 6\Delta \}$ and $\mathcal{D}_t \triangleq \{\forall \tau \leq t : (\alpha_{1,t}^2, \alpha_{2,t}^2) \in \mathcal{D} \}$. Note that $(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 < 6\Delta$ 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

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

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

$$\tilde{d}_{t+1} \triangleq |H_{t+1} \mathbf{1}_{\mathcal{D}_t} - E[H_{t+1} \mathbf{1}_{\mathcal{D}_t} | \mathcal{F}_t]| \leq (1 - 4\eta \sigma^2)^{-t-1}(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 - E[(\alpha_{1,t+1}^2 - \alpha_{2,t+1}^2)^2 | \mathcal{F}_t]| \mathbf{1}_{\mathcal{D}_t} \leq (1 - 4\eta \sigma^2)^{-t-1}\tilde{C}_5,$$

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

$$P \left( H_t \mathbf{1}_{\mathcal{D}_{t-1}} - H_0 \geq O(1)\tilde{r}_t \log^\frac{1}{\eta^2} \sqrt{\Delta} \right) \leq \exp \left( -O(1)\tilde{r}_t^2 \log \left( \frac{1}{\eta^2} \right) \right) = O(\eta^2 \delta).$$

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

$$((\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 - \Delta) \mathbf{1}_{\mathcal{D}_{t-1}} < (1 - 4\eta \sigma^2)^t((\alpha_{1,0}^2 - \alpha_{2,0}^2)^2 + O(1)\tilde{r}_t \log^\frac{1}{\eta^2} \sqrt{\Delta})$$

$$< (\alpha_{1,0}^2 - \alpha_{2,0}^2)^2 + O((1 - 4\eta \sigma^2)^t\tilde{r}_t \log^\frac{1}{\eta^2} \sqrt{\Delta}) < 4\Delta + O \left( \frac{\sqrt{\Delta \eta}}{\sigma^2} \left( \sqrt{\log \frac{d_1 + d_2}{\eta^2}} + \sqrt{\log \frac{1}{\delta}} \right) \log^\frac{1}{\eta^2} \right).$$
When $D_{t-1}$ holds, w.p. at least $1 - O(\eta^2 \delta)$, by the inequality above we have
\[
(\alpha_{1,t}^2 - \alpha_{2,t}^2)^2 < 5\Delta + O\left(\frac{\sqrt{\Delta \eta}}{\sigma^2} \left(\sqrt{\log \frac{d_1 + d_2}{\eta^2}} + \sqrt{\log \frac{1}{\delta}}\right) \log^2 \left(\frac{1}{\eta^2 \delta}\right)\right) < 6\Delta.
\]
Note that to make sure last terms is smaller than $\Delta$, we need
\[
\frac{\sqrt{\eta}}{\sigma^2 \sqrt{\Delta}} \left(\sqrt{\log \frac{d_1 + d_2}{\eta^2}} + \sqrt{\log \frac{1}{\delta}}\right) \log^2 \left(\frac{1}{\eta^2 \delta}\right) = O(1).
\]
As $\Delta = O\left(\frac{\log \frac{1}{\eta^2 \delta}}{\log \frac{1}{\delta}}\right)$, we know that as long as $\eta$ is polynomial in $\sigma$, choosing $\sigma = O((\log \frac{1}{\delta})^{-\frac{1}{3}})$ is sufficient. Now we know that if $D_{t-1}$ holds, $D_t$ 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 $O(\frac{1}{\eta^2 \delta})$ steps w.p. at least $1 - \delta$.

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

**Proof.**
\[
|1 - x^T_i u_*| < (1 + x^T_i u_*)|1 - x^T_i u_*| \\
= |1 - (x^T_i u_*)^2 + x^T_i u_* v^T_i y_t - x^T_i u_* v^T_i y_t| \\
\leq |1 - x^T_i u_* v^T_i y_t| + |(x^T_i u_*)^2 - x^T_i u_* v^T_i y_t| \\
= |1 - x^T_i M v_t| + |x^T_i u_* - v^T_i y_t| \\
\leq |1 - x^T_i M v_t| + \sqrt{\frac{2}{\sigma}} |x^T_i u_* - v^T_i y_t| \\
< \sqrt{6\gamma} + \frac{\sqrt{2}}{\sigma} \sqrt{6\Delta}.
\]

The last inequality comes from Lemmas 6 and 9. Together with Lemma 1 we can get
\[
\|x_t - u_*\|^2 = (1 - x^T_i u_*)^2 + \|\beta_{1,t}\|^2 < (\sqrt{6\gamma} + \frac{\sqrt{2}}{\sigma} \sqrt{6\Delta})^2 + 2 \frac{\eta C_2}{\sigma^2} = O(\sigma^2 + \frac{\eta}{\sigma^{10}}).
\]
Note that we use $C_2 = O(\frac{1}{\sqrt{\sigma^3}})$ 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 $\eta \leq \eta_5 = O(\sigma^{10} \epsilon)$, we will have $\|x_t - u_*\|^2 < \epsilon$ and $\|y_t - v_*\|^2 < \epsilon$.

## 4 Proof of Theorem 2

Recall that the gradient of $\tilde{F}$ takes the following form.
\[
\nabla_X \tilde{F}(X, Y) = (XY^T - M)Y - d_2 \sigma_2^2 X, \\
\nabla_Y \tilde{F}(X, Y) = (XY^T - M)^T X - d_1 \sigma_1^2 Y.
\]

Suppose $(U, V)$ is a stationary point. Then we have
\[
(UV^T - M)V - d_2 \sigma_2^2 U = 0, \tag{13}
\]
\[
(UV^T - M)^T U - d_1 \sigma_1^2 V = 0. \tag{14}
\]

**Step 1:** To prove the first statement, simply left multiply each side of (13) by $U^T$ and each side of (14) by $V^T$, and we have the following equations.
\[
U^TUV^T V - UV^T MV - d_2 \sigma_2^2 U^T U = 0, \\
V^T VU^T U - V^T M^T U - d_1 \sigma_1^2 V^T V = 0.
\]
Note that the following equation naturally holds.
\[ U^\top UV^\top V - U^\top MV = (V^\top VU^\top U - V^\top M^\top U)^\top. \]
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 \((\bar{U}, \bar{V})R\) is a stationary point, where
\[(\bar{U}, \bar{V}) = (\sqrt{\gamma}A(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}, \frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}), \]
where \(R \in \mathbb{R}^{r \times r}\) is an orthogonal matrix. We only need to check (13) and (14). In fact, we have
\[ \nabla X \bar{F}(UR, VR) = -\gamma \sigma^2 AB^\top \frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2 \gamma^2 B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R \]
\[ = -\sigma^2 \gamma \sqrt{\gamma}B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2 \gamma^2 B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R = 0, \]
and
\[ \nabla Y \bar{F}(UR, VR) = -\gamma \sigma^2 BA^\top \frac{1}{\sqrt{\gamma}}B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2 \gamma^2 B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R \]
\[ = -\sigma^2 \gamma \sqrt{\gamma}B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R + \sigma^2 \gamma^2 B(\Sigma - \gamma \sigma^2 I_r)^{\frac{1}{2}}R = 0. \]
Combine the above equations together and we know that \((\bar{U}, \bar{V})\) is a stationary point.

**Step 3:** We next show that \(\{(\bar{U}, \bar{V})R\}|R \in \mathbb{R}^{r \times r}, \text{orthogonal}\) 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 \(\nabla^2 \bar{F}(X, Y)\). 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
\[ [\nabla^2 \bar{F}(W)](Z_1, Z_2) = \sum_{i,j,k,l} \frac{\partial^2 \bar{F}(W)}{\partial W[i,j] \partial W[k,l]} Z_1[i,j]Z_2[k,l], \forall Z_1, Z_2 \in \mathbb{R}^{(d_1 + d_2) \times r}. \]
We can then express the Hessian \(\nabla^2 \bar{F}(W)\) as follows:
\[ [\nabla^2 \bar{F}(X, Y)](\Delta, \Delta) = 2 < XY^\top - M, \Delta_U \Delta_V^\top > + \|U \Delta_V^\top + \Delta_U V^\top\|^2_F + \sigma^2 \|\Delta_U\|^2_F + \sigma^2 \|\Delta_V\|^2_F, \]
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 \(W = \begin{bmatrix} X \\ Y \end{bmatrix}, \bar{W} = \begin{bmatrix} \bar{U} \\ \bar{V} \end{bmatrix}\), and \(\bar{M} = \bar{U} \bar{V}^\top\),
\[ R = \arg\min_{R \in \mathbb{R}^{r \times r}, \text{orthogonal}} \|W - \bar{W} R\|. \]
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 \(\nabla \bar{F}(U, V) = 0\), we denote \(\Delta = \begin{bmatrix} U - \bar{U} R \\ V - \bar{V} R \end{bmatrix}\), then we have
\[ [\nabla^2 \bar{F}(U, V)](\Delta, \Delta) \leq -\|U V^\top - \bar{M}\|^2_F - 3\sigma^2 \|A^\top U - B^\top V\|^2_F. \]
Moreover, \([\nabla^2 \bar{F}(U, V)](\Delta, \Delta) < 0\) if
\[ (U, V) \notin \{(\bar{U}, \bar{V})R | R \in \mathbb{R}^{r \times r}, R^\top R = R^\top R = I_r\}. \]
Proof. Recall that the quadratic form defined by the Hessian can be written as follows.

\[ [\nabla^2 \bar{F}(U, V)](\Delta, \Delta) = 2 < UV^T - M, \Delta_U \Delta_V^\top > + \| \Delta_U \Delta_V^\top \|^2_F + \sigma^2 \| \Delta_U \|^2_F + \sigma^2 \| \Delta_V \|^2_F. \]  

(16)

We start from the second term \( \| \Delta_U \Delta_V^\top + \Delta_U V^T \|^2_F \). Similar to the proof of Claim B.5 in Du et al. (2018), we have

\[
\| \Delta_U \Delta_V^\top + \Delta_U V^T \|^2_F = \| \Delta_U \Delta_V^\top + UV^T - \tilde{M} \|^2_F \\
= \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F + 2 < \Delta_U \Delta_V^\top, UV^T - \tilde{M} > \\
= \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F + 2 < \Delta_U \Delta_V^\top, UV^T - M > + 2 < \Delta_U \Delta_V^\top, M - \tilde{M} > \\
= \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - M \|^2_F + 2 < \Delta_U \Delta_V^\top, UV^T - M > + 2\sigma^2 < \Delta_U \Delta_V^\top, AB^T > \\

\]

Plugging this equation into (16), we have

\[
[\nabla^2 \bar{F}(U, V)](\Delta, \Delta) = 4 < UV^T - M, \Delta_U \Delta_V^\top > + \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F \\
\quad + \sigma^2 (\| \Delta_U \|^2_F + \| \Delta_V \|^2_F + 2 < \Delta_U \Delta_V^\top, AB^T > ).
\]

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

\[
4 < UV^T - M, \Delta_U \Delta_V^\top > = 4 < UV^T - M, \tilde{M} > - 2\sigma^2 (\| \Delta_U \|^2_F + \| \Delta_V \|^2_F ) + 2\sigma^2 (\| \tilde{U} \|^2_F + \| \tilde{V} \|^2_F ) \\
= -4\| UV^T - \tilde{M} \|^2_F - 4\sigma^2 < AB^T, \tilde{M} - UV^T > \\
\quad + 2\sigma^2 (\| \tilde{U} \|^2_F + \| \tilde{V} \|^2_F - \| U \|^2_F - \| V \|^2_F ).
\]

Thus,

\[
[\nabla^2 \bar{F}(U, V)](\Delta, \Delta) = -4\| UV^T - \tilde{M} \|^2_F + \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F \\
\quad - \sigma^2 \left[ \| \Delta_U \|^2_F + \| \Delta_V \|^2_F - 2 < \Delta_U \Delta_V^\top, AB^T > \\
\quad - 2 \left( \| \tilde{U} \|^2_F + \| \tilde{V} \|^2_F - \| U \|^2_F - \| V \|^2_F \right) + 4 < AB^T, \tilde{M} - UV^T > \right].
\]

(17)

We then have the following two claims:

**Claim 1.** \(-4\| UV^T - \tilde{M} \|^2_F + \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F \leq -\| UV^T - \tilde{M} \|^2_F.\)

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

\[
\| \Delta_U \Delta_V^\top \|^2_F \leq \frac{1}{4} \| \Delta \|^2 \| \Delta \|^2_F \\
\leq \frac{1}{2} \| WW^T - \tilde{W} \tilde{W}^T \|^2_F \\
= 2\| UV^T - \tilde{M} \|^2_F - \| U^T \tilde{U} - V^T \tilde{V} \|^2_F + \frac{1}{2} \| U^T \tilde{U} - V^T \tilde{V} \|^2_F + \frac{1}{2} \| \tilde{U}^T \tilde{U} - \tilde{V}^T \tilde{V} \|^2_F \\
= 2\| UV^T - \tilde{M} \|^2_F - \| U^T \tilde{U} - V^T \tilde{V} \|^2_F \\
\leq 2\| UV^T - \tilde{M} \|^2_F.
\]

Thus,

\[
-4\| UV^T - \tilde{M} \|^2_F + \| \Delta_U \Delta_V^\top \|^2_F + \| UV^T - \tilde{M} \|^2_F \leq -4\| UV^T - \tilde{M} \|^2_F + 2\| UV^T - \tilde{M} \|^2_F + \| UV^T - \tilde{M} \|^2_F \\
= -\| UV^T - \tilde{M} \|^2_F.
\]

\[\square\]

**Claim 2.** \(\| U \|^2_F + \| V \|^2_F - \| \tilde{U} \|^2_F - \| \tilde{V} \|^2_F + 2 < AB^T, \tilde{M} - UV^T > = \| A^T \Delta_U - B^T \Delta_V \|^2_F.\)
Proof. First, the LHS of the equation can be rewritten as follows.

\[
\|U\|^2_F + \|V\|^2_F - \|\tilde{U}\|^2_F - \|\tilde{V}\|^2_F + 2 < AB^T, \tilde{M} - UV^T > \\
= \|U\|^2_F + \|V\|^2_F - 2 < AB^T, \tilde{M} + UV^T > + 4 < AB^T, \tilde{M} > - \|\tilde{U}\|^2_F - \|\tilde{V}\|^2_F \\
= \|U\|^2_F + \|V\|^2_F - 2 < AB^T, \tilde{M} + UV^T > + 4 \|\tilde{U}\|^2_F - \|\tilde{V}\|^2_F \\
= \|U\|^2_F + \|V\|^2_F + \|\tilde{U}\|^2_F + \|\tilde{V}\|^2_F - 2 < AB^T, \tilde{M} + UV^T > ,
\]

where we use the fact

\[
< AB^T, \tilde{M} >= \|\tilde{U}\|^2_F = \|\tilde{V}\|^2_F .
\]

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

\[
\|A^T \Delta_U - B^T \Delta_V\|^2_F = \|\Delta_U\|^2_F + \|\Delta_V\|^2_F - 2 < \Delta_U \Delta_V^T, AB^T > \\
= \|U\|^2_F + \|V\|^2_F + \|\tilde{U}\|^2_F + \|\tilde{V}\|^2_F - 2 < AB^T, \tilde{M} + UV^T > \\
- 2 \text{tr}(U(\tilde{U}R)^T - 2 \text{tr}(V(\tilde{V}R)^T) + 2 \text{tr}(U(\tilde{U}R)^T) + 2 \text{tr}(V(\tilde{V}R)^T)) \\
= \|U\|^2_F + \|V\|^2_F + \|\tilde{U}\|^2_F + \|\tilde{V}\|^2_F - 2 < AB^T, \tilde{M} + UV^T > \\
= \|U\|^2_F + \|V\|^2_F - \|\tilde{U}\|^2_F - \|\tilde{V}\|^2_F + 2 < AB^T, \tilde{M} - UV^T > .
\]

\[
\n\]

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

\[
[\nabla^2 \mathcal{F}(U,V)](\Delta, \Delta) \leq -\|UV^T - \tilde{M}\|^2_F - 3\sigma^2\|A^T \Delta_U - B^T \Delta_V\|^2_F .
\]

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

Claim 3. \(\|UV^T - \tilde{M}\|^2_F + 3\sigma^2\|A^T U - B^T V\|^2_F = 0\) if and only if

\[
(U,V) = (\tilde{U}, \tilde{V})R ,
\]

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

Proof. \(\|UV^T - \tilde{M}\|^2_F + 3\sigma^2\|A^T U - B^T V\|^2_F = 0\) if and only if

\[
UV^T = \tilde{M}, \quad A^T U = B^T V .
\]

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

\[
A^T UV^T = (\Sigma - \sigma^2 I)B^T \iff B^T VV^T = (\Sigma - \sigma^2 I)B^T \\
\iff VV^T = B(\Sigma - \sigma^2 I)B^T \\
\iff V = B(\Sigma - \sigma^2 I)^{-\frac{1}{2}} R = \tilde{V}R'.
\]

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

As a direct result of Claim 3, for stationary point \((U,V) \notin \{(\tilde{U}, \tilde{V})R' | R' \in \mathbb{R}^{r \times r}, R'R'^T = R'^TR' = I_r\}\), we have

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

We prove the lemma.

Lemma 10 directly implies that \(\{(\tilde{U}, \tilde{V})R' | R' \in \mathbb{R}^{r \times r}, R'R'^T = R'^TR' = I_r\}\) 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 $\eta$, noise level $\sigma_1, \sigma_2$, matrix $M \in \mathbb{R}^{d_1 \times d_2}$, number of iterations $T$.

**Initialize:** initialize $(x_0, y_0)$ arbitrarily.

for $t = 0 \ldots T - 1$ do
    Sample $\xi_{1,t} \sim N(0, \sigma_1^2 I_{d_1})$ and $\xi_{2,t} \sim N(0, \sigma_2^2 I_{d_2})$.
    $\bar{x}_t = x_t + \xi_{1,t}, \quad \bar{y}_t = y_t + \xi_{2,t}$.
    $x_{t+1} = x_t - \eta(\bar{x}_t \bar{y}^\top_t - M)\bar{y}_t$.
    $y_{t+1} = y_t - \eta(\bar{y}_t \bar{x}^\top_t - M^\top)\bar{x}_t$.
end for

6 Figures

Figure 1: The visualization of objective functions $\mathcal{F}(x, y) = (1 - xy)^2$ and $\tilde{\mathcal{F}}(x, y)$ 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)). $\tilde{\mathcal{F}}(x, y)$ 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 30~40 iterations, as shown in (a,b,c). GD does not show phase transitions.
Figure 5: The visualization of objective $F(x, y) = (1 - xy)^2$. 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 $2 \times 10^3$ iterations.
References
