Algorithm 4: Find a path $\{W_{\mu_k}\}$ via a particular scheduling for μ_k when a is unknown.

Input: $\mu_0 \in \left[\frac{a^2}{4(a^2+1)^3}, \frac{a^2}{4}\right), \varepsilon > 0$

Output: $\{W_{\mu_k}\}_{k=0}^{\infty}$

$\widehat{a} \leftarrow \sqrt{4(\mu_0 + \varepsilon)}$ // $\forall \varepsilon \geq 0$ `s.t.` $\widehat{a} < a$

$W_{\mu_0} \leftarrow \texttt{GradientFlow}(g_{\mu_0}, \mathbf{0})$

for $k = 1, 2, \ldots$ **do**

  Let $\mu_{k+1} \in \left[(2/\widehat{a})^{2/3} \mu_k^{4/3}, \mu_k\right)$

  $W_{\mu_{k+1}} \leftarrow \texttt{GradientFlow}(g_{\mu_{k+1}}, W_{\mu_k})$

end

return $\{W_{\mu_k}\}_{k=0}^{\infty}$

A Practical Implementation of Algorithm 2

We present a practical implementation of our homotopy algorithm in Algorithm 4. The updating scheme for μ_k is now independent of the parameter a, but as presented, the initialization for μ_0 still depends on a. This is for the following reason: It is possible to make the updating scheme independent of a *without imposing any additional assumptions on* a, as evidenced by Lemma 4 below. The initialization for μ_0, however, is trickier, and we must consider two separate cases:

1. *No assumptions on* a. In this case, if a is too small, then the problem becomes harder and the initial choice of μ_0 matters.
2. *Lower bound on* a. If we are willing to accept a lower bound on a, then there is an initialization for μ_0 that does not depend on a.

In Corollary 1, we illustrate this last point with the additional condition that $a > \sqrt{5/27}$. This essentially amounts to an assumption on the minimum signal, and is quite standard in the literature on learning SEM.

Lemma 4. *Under the assumption* $\frac{a^2}{4(a^2+1)^3} \leq \mu_0 < \frac{a^2}{4}$, *the Algorithm 4 outputs the global optimal solution to (6), i.e.*

$$\lim_{k \to \infty} W_{\mu_k} = W_{\mathsf{G}}.$$

It turns out that the assumption in Lemma 4 is not overly restrictive, as there exist pre-determined sequences of $\{\mu_k\}_{k=0}^{\infty}$ that can ensure the effectiveness of Algorithm 4 for any values of a greater than a certain threshold.

B From Population Loss to Empirical Loss

The transformation from population loss to empirical can be thought from two components. First, with a given empirical loss, Algorithms 2 and 3 still achieve the global minimum, W_{G}, of problem 6, but now the output from the Algorithm is an empirical estimator $\hat{a}$, rather than ground truth a, Theorem 1 and Corollary 1 would continue to be valid. Second, the global optimum, W_{G}, of the empirical loss possess the same DAG structure as the underlying W_*. The finite-sample findings in Section 5 (specifically, Lemmas 18 and 19) of Loh and Bühlmann [31], which offer sufficient conditions on the sample size to ensure that the DAG structures of W_{G} and W_* are identical.

C From Continuous to Discrete: Gradient Descent

Previously, gradient flow was employed to address the intermediate problem (7), a method that poses implementation challenges in a computational setting. In this section, we introduce Algorithm 6 that leverages gradient descent to solve (7) in each iteration. This adjustment serves practical considerations. We start with the convergence results of Gradient Descent.

Definition 1. *f is L-smooth, if f is differentiable and $\forall x, y \in dom(f)$ such that* $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$.

Algorithm 5: Gradient Descent(f, η, W_0, ϵ)

Input: function f, step size η, initial point W_0, tolerance ϵ
Output: W_t

$t \leftarrow 0$
while $\|\nabla f(W_t)\|_2 > \epsilon$ **do**
  $W_{t+1} \leftarrow W_t - \eta \nabla f(W_t)$
  $t \leftarrow t + 1$
end

Algorithm 6: Homotopy algorithm using gradient descent for solving (1).

Input: Initial $W_{-1} = W(x_{-1}, y_{-1})$, $\mu_0 \in \left[\frac{a^2}{4(a^2+1)^3}\frac{(1+\beta)^4}{(1-\beta)^2}, \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}\right)$,
$\eta_0 = \frac{1}{\mu_0(a^2+1)+3a^2}$, $\epsilon_0 = \min\{\beta a \mu_0, \mu_0^{3/2}\}$
Output: $\{W_{\mu_k}\}_{k=0}^{\infty}$

$W_{\mu_0,\epsilon_0} \leftarrow$ `Gradient Descent`$(g_{\mu_0}, \eta_0, W_{-1}, \epsilon_0)$
for $k = 1, 2, \ldots$ **do**
  Let $\mu_k = (2\mu_{k-1}^2)^{2/3}\frac{(a+\epsilon_{k-1}/\mu_{k-1})^{2/3}}{(a-\epsilon_{k-1}/\mu_{k-1})^{4/3}}$
  Let $\eta_k = \frac{1}{\mu_k(a^2+1)+3a^2}$
  Let $\epsilon_k = \min\{\beta a \mu_k, \mu_k^{3/2}\}$
  $W_{\mu_k,\epsilon_k} \leftarrow$ `Gradient Descent`$(g_{\mu_k}, \eta_k, W_{\mu_{k-1}}, \epsilon_k)$
end

Theorem 3 (Nesterov et al. 33)**.** *If function f is L-smooth, then Gradient Descent (Algorithm 5) with step size $\eta = 1/L$, finds an ϵ-first-order stationary point (i.e. $\|\nabla f(x)\|_2 \leq \epsilon$) in $2L(f(x^0) - f^*)/\epsilon^2$ iterations.*

One of the pivotal factors influencing the convergence of gradient descent is the selection of the step size. Theorem 3 select a step size $\eta = \frac{1}{L}$. Therefore, our initial step is to determine the smoothness of $g_\mu(W)$ within our region of interest, $A = \{0 \leq x \leq a, 0 \leq y \leq \frac{a}{a^2+1}\}$.

Lemma 5. *Consider the function $g_\mu(W)$ as defined in Equation 7 within the region $A = \{0 \leq x \leq a, 0 \leq y \leq \frac{a}{a^2+1}\}$. It follows that for all $\mu \geq 0$, the function $g_\mu(W)$ is $\mu(a^2+1) + 3a^2$-smooth.*

Since gradient descent is limited to identifying the ϵ stationary point of the function. Thus, we study the gradient of $g_\mu(W) = \mu f(W) + h(W)$, i.e. $\nabla g_\mu(W)$ has the following form

$$\nabla g_\mu(W) = \begin{pmatrix} \mu(x-a) + y^2 x \\ \mu(a^2+1)y - a\mu + yx^2 \end{pmatrix}$$

As gradient descent is limited to identifying the ϵ stationary point of the function, we, therefore, focus on $\|g_\mu(W)\|_2 \leq \epsilon$. This can be expressed in the subsequent manner:

$$\|\nabla g_\mu(W)\|_2 \leq \epsilon \Rightarrow -\epsilon \leq \mu(x-a) + y^2 x < \epsilon \quad \text{and} \quad -\epsilon \leq \mu(a^2+1)y - a\mu + yx^2 \leq \epsilon$$

As a result,

$$\{(x,y) \mid \|\nabla g_\mu(W)\|_2 \leq \epsilon\} \subseteq \{(x,y) \mid \frac{\mu a - \epsilon}{\mu + y^2} \leq x \leq \frac{\mu a + \epsilon}{\mu + y^2}, \frac{\mu a - \epsilon}{x^2 + \mu(a^2+1)} \leq y \leq \frac{\mu a + \epsilon}{x^2 + \mu(a^2+1)}\}$$

Here we denote such region as $A_{\mu,\epsilon}$

$$A_{\mu,\epsilon} = \{(x,y) \mid \frac{\mu a - \epsilon}{\mu + y^2} \leq x \leq \frac{\mu a + \epsilon}{\mu + y^2}, \frac{\mu a - \epsilon}{x^2 + \mu(a^2+1)} \leq y \leq \frac{\mu a + \epsilon}{x^2 + \mu(a^2+1)}\} \tag{10}$$

Figure 6 and 7 illustrate the region $A_{\mu,\epsilon}$.

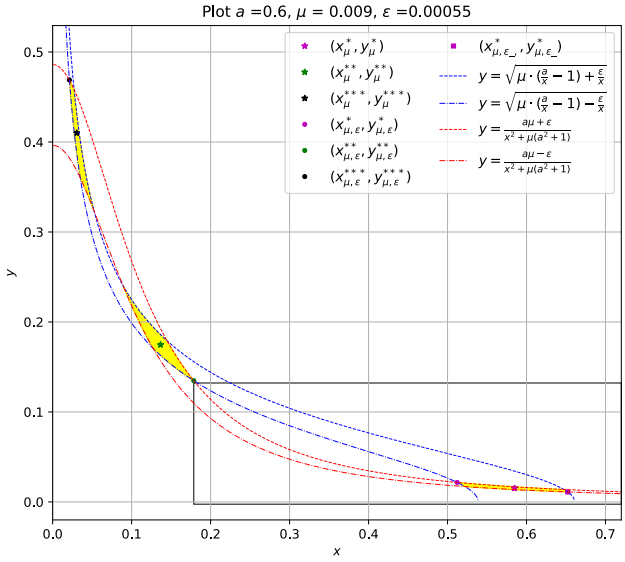


Figure 6: An example of $A_{\mu,\epsilon}$ is depicted for $a = 0.6$, $\mu = 0.009$, and $\epsilon = 0.00055$. The yellow region signifies ϵ stationary points, denoted as $A_{\mu,\epsilon}$ and defined by Equation (10). $A_{\mu,\epsilon}$ is the disjoint union of $A^1_{\mu,\epsilon}$ and $A^2_{\mu,\epsilon}$, which are defined by Equations (21) and (22), respectively.

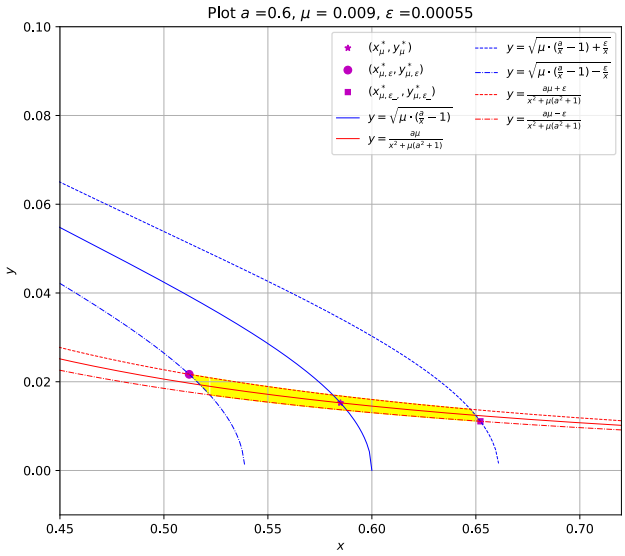


Figure 7: Here is a localized illustration of $A_{\mu,\epsilon}$ that includes the point (x^*_μ, y^*_μ). This region, referred to as $A^1_{\mu,\epsilon}$, is defined in Equation (21).

Given that the gradient descent can only locate ϵ stationary points within the region $A_{\mu,\epsilon}$ during each iteration, the boundary of $A_{\mu,\epsilon}$ becomes a critical component of our analysis. To facilitate clear presentation, it is essential to establish some pertinent notations.

- $$\begin{cases} x = \dfrac{\mu a}{\mu + y^2} & \text{(11a)} \\ y = \dfrac{\mu a}{\mu(a^2+1) + x^2} & \text{(11b)} \end{cases}$$

  If the system of equations yields only a single solution, we denote this solution as (x_μ^*, y_μ^*). If it yields two solutions, these solutions are denoted as $(x_\mu^*, y_\mu^*), (x_\mu^{**}, y_\mu^{**})$, with $x_\mu^{**} < x_\mu^*$. In the event that there are three distinct solutions to the system of equations, these solutions are denoted as $(x_\mu^*, y_\mu^*), (x_\mu^{**}, y_\mu^{**}), (x_\mu^{***}, y_\mu^{***})$, where $x_\mu^{***} < x_\mu^{**} < x_\mu^*$.

- $$\begin{cases} x = \dfrac{\mu a - \epsilon}{\mu + y^2} & \text{(12a)} \\ y = \dfrac{\mu a + \epsilon}{\mu(a^2+1) + x^2} & \text{(12b)} \end{cases}$$

  If the system of equations yields only a single solution, we denote this solution as $(x_{\mu,\epsilon}^*, y_{\mu,\epsilon}^*)$. If it yields two solutions, these solutions are denoted as $(x_{\mu,\epsilon}^*, y_{\mu,\epsilon}^*), (x_{\mu,\epsilon}^{**}, y_{\mu,\epsilon}^{**})$, with $x_{\mu,\epsilon}^{**} < x_{\mu,\epsilon}^*$. In the event that there are three distinct solutions to the system of equations, these solutions are denoted as $(x_{\mu,\epsilon}^*, y_{\mu,\epsilon}^*), (x_{\mu,\epsilon}^{**}, y_{\mu,\epsilon}^{**}), (x_{\mu,\epsilon}^{***}, y_{\mu,\epsilon}^{***})$, where $x_{\mu,\epsilon}^{***} < x_{\mu,\epsilon}^{**} < x_{\mu,\epsilon}^*$.

- $$\begin{cases} x = \dfrac{\mu a + \epsilon}{\mu + y^2} & \text{(13a)} \\ y = \dfrac{\mu a - \epsilon}{\mu(a^2+1) + x^2} & \text{(13b)} \end{cases}$$

  If the system of equations yields only a single solution, we denote this solution as $(x_{\mu,\epsilon\_}^*, y_{\mu,\epsilon\_}^*)$. If it yields two solutions, these solutions are denoted as $(x_{\mu,\epsilon\_}^*, y_{\mu,\epsilon\_}^*), (x_{\mu,\epsilon\_}^{**}, y_{\mu,\epsilon\_}^{**})$, with $x_{\mu,\epsilon\_}^{**} < x_{\mu,\epsilon\_}^*$. In the event that there are three distinct solutions to the system of equations, these solutions are denoted as $(x_{\mu,\epsilon\_}^*, y_{\mu,\epsilon\_}^*), (x_{\mu,\epsilon\_}^{**}, y_{\mu,\epsilon\_}^{**}), (x_{\mu,\epsilon\_}^{***}, y_{\mu,\epsilon\_}^{***})$, where $x_{\mu,\epsilon\_}^{***} < x_{\mu,\epsilon\_}^{**} < x_{\mu,\epsilon\_}^*$.

Remark 4. *There always exists at least one solution to the above system of equations. When μ is sufficiently small, the above system of equations always yields three solutions, as demonstrated in Theorem 5 and Theorem 9.*

The parameter ϵ can substantially influence the behavior of the systems of equations (12a),(12b) and (13a),(13b). A crucial consideration is to ensure that ϵ remains adequately small. To facilitate this, we introduce a new parameter, β, whose specific value will be determined later. At this stage, we merely require that β should lie within the interval $(0, 1)$. We further impose a constraint on ϵ to satisfy the following inequality:

$$\epsilon \leq \beta a \mu \tag{14}$$

Following the same procedure when we deal with $\epsilon = 0$. Let us substitute (12a) into (12b), then we obtain an equation that only involves the variable y

$$r_\epsilon(y;\mu) = \frac{a + \epsilon/\mu}{y} - (a^2+1) - \frac{(\mu a - \epsilon)^2/\mu}{(y^2+\mu)^2} \tag{15}$$

Let us substitute (12b) into (12a), then we obtain an equation that only involves the variable x

$$t_\epsilon(x;\mu) = \frac{a - \epsilon/\mu}{x} - 1 - \frac{(\mu a + \epsilon)^2/\mu}{(\mu(a^2+1) + x^2)^2} \tag{16}$$

Proceed similarly for equations (13a) and (13b).

$$r_{\epsilon\_}(y;\mu) = \frac{a - \epsilon/\mu}{y} - (a^2+1) - \frac{(\mu a + \epsilon)^2/\mu}{(y^2+\mu)^2} \tag{17}$$

$$t_{\epsilon\_}(x;\mu) = \frac{a+\epsilon/\mu}{x} - 1 - \frac{(\mu a-\epsilon)^2/\mu}{(\mu(a^2+1)+x^2)^2} \tag{18}$$

Given the substantial role that the system of equations 12a and 12b play in our analysis, the existence of ϵ in these equations complicates the analysis, this can be avoided by considering the worst-case scenario, i.e., when $\epsilon = \beta a\mu$. With this particular choice of ϵ, we can reformulate (15) and (16) as follows, denoting them as $r_\beta(y;\epsilon)$ and $r_\beta(x;\epsilon)$ respectively.

$$r_\beta(y;\mu) = \frac{a(1+\beta)}{y} - (a^2+1) - \frac{\mu a^2(1-\beta)^2}{(y^2+\mu)^2} \tag{19}$$

$$t_\beta(x;\mu) = \frac{a(1-\beta)}{x} - 1 - \frac{\mu a^2(1+\beta)^2}{(\mu(a^2+1)+x^2)^2} \tag{20}$$

The functions $r_\epsilon(y;\mu)$, $r_{\epsilon\_}(y;\mu)$, and $r_\beta(y;\mu)$ possess similar properties to $r(y;\mu)$ as defined in Equation (8), with more details available in Theorem 7 and 8. Additionally, the functions $t_\epsilon(x;\mu)$, $t_{\epsilon\_}(x;\mu)$, and $t_\beta(x;\mu)$ share similar characteristics with $t(x;\mu)$ as defined in Equation (9), with more details provided in Theorem 9.

As illustrated in Figure 6, the ϵ-stationary point region $A_{\mu,\epsilon}$ can be partitioned into two distinct areas, of which only the lower-right one contains (x^*_μ, y^*_μ) and it is of interest to our analysis. Moreover, $(x^*_{\mu,\epsilon}, y^*_{\mu,\epsilon})$ and $(x^{**}_{\mu,\epsilon}, y^{**}_{\mu,\epsilon})$ are extremal point of two distinct regions. The upcoming corollary substantiates this intuition.

Corollary 3. *If $\mu < \tau$ (τ is defined in Theorem 5(v)), assume ϵ satisfies (14), β satisfies $\left(\frac{1+\beta}{1-\beta}\right)^2 \le a^2+1$, systems of equations (12a),(12b) at least have two solutions. Moreover, $A_{\mu,\epsilon} = A^1_{\mu,\epsilon} \cup A^2_{\mu,\epsilon}$*

$$A^1_{\mu,\epsilon} = A_{\mu,\epsilon} \cap \{(x,y) \mid x \ge x^*_{\mu,\epsilon}, y \le y^*_{\mu,\epsilon}\} \tag{21}$$

$$A^2_{\mu,\epsilon} = A_{\mu,\epsilon} \cap \{(x,y) \mid x \le x^{**}_{\mu,\epsilon}, y \ge y^{**}_{\mu,\epsilon}\} \tag{22}$$

Corollary 3 suggests that $A_{\mu,\epsilon}$ can be partitioned into two distinct regions, namely $A^1_{\mu,\epsilon}$ and $A^2_{\mu,\epsilon}$. Furthermore, for every (x,y) belonging to $A^1_{\mu,\epsilon}$, it follows that $x \ge x^*_{\mu,\epsilon}$ and $y \le y^*_{\mu,\epsilon}$. Similarly, for every (x,y) that lies within $A^2_{\mu,\epsilon}$, the condition $x \le x^{**}_{\mu,\epsilon}$ and $y \ge y^{**}_{\mu,\epsilon}$ holds. The region $A^1_{\mu,\epsilon}$ represents the "correct" region that gradient descent should identify. In this context, identifying the region equates to pinpointing the extremal points of the region. As a result, our focus should be on the extremal points of $A^1_{\mu,\epsilon}$ and $A^2_{\mu,\epsilon}$, specifically at $(x^*_{\mu,\epsilon}, y^*_{\mu,\epsilon})$ and $(x^{**}_{\mu,\epsilon}, y^{**}_{\mu,\epsilon})$. Furthermore, the key to ensuring the convergence of the gradient descent to the $A^1_{\mu,\epsilon}$ is to accurately identify the "basin of attraction" of the region $A^1_{\mu,\epsilon}$. The following lemma provides a region within which, regardless of the initialization point of the gradient descent, it converges inside $A^1_{\mu,\epsilon}$.

Lemma 6. *Assume $\mu < \tau$ (τ is defined in Theorem 5(v)), $\left(\frac{1+\beta}{1-\beta}\right)^2 \le a^2+1$. Define $B_{\mu,\epsilon} = \{(x,y) \mid x^{**}_{\mu,\epsilon} < x \le a, 0 \le y < y^{**}_{\mu,\epsilon}\}$. Run Algorithm 5 with input $f = g_\mu(x,y), \eta = \frac{1}{\mu(a^2+1)+3a^2}, W_0 = (x(0), y(0))$, where $(x(0), y(0)) \in B_{\mu,\epsilon}$, then after at most $\frac{2(\mu(a^2+1)+3a^2)(g_\mu(x(0),y(0))-g_\mu(x^*_\mu,y^*_\mu))}{\epsilon^2}$ iterations, $(x_t, y_t) \in A^1_{\mu,\epsilon}$.*

Lemma 6 can be considered the gradient descent analogue of Lemma 2. It plays a pivotal role in the proof of Theorem 4. In Figure 6, the lower-right rectangle corresponds to $B_{\mu,\epsilon}$. Lemma 6 implies that the gradient descent with any initialization inside $B_{\mu_{k+1},\epsilon_{k+1}}$ will converge to $A^1_{\mu_{k+1},\epsilon_{k+1}}$ at last. Then, by utilizing the previous solution W_{μ_k,ϵ_k} as the initial point, as long as it lies within region $B_{\mu_{k+1},\epsilon_{k+1}}$, the gradient descent can converge to $A^1_{\mu_{k+1},\epsilon_{k+1}}$ which is ϵ stationary points region that contains $W^*_{\mu_{k+1}}$, thereby achieving the goal of tracking $W^*_{\mu_{k+1}}$. Following the scheduling for μ_k prescribed in Algorithm 6 provides a sufficient condition to ensure that will happen.

We now proceed to present the theorem which guarantees the global convergence of Algorithm 6.

Theorem 4. *If* $\delta \in (0,1)$, $\beta \in (0,1)$, $\left(\frac{1+\beta}{1-\beta}\right)^2 \le (1-\delta)(a^2+1)$, *and* μ_0 *satisfies*

$$\frac{a^2}{4(a^2+1)^3} \le \frac{a^2}{4(a^2+1)^3}\frac{(1+\beta)^4}{(1-\beta)^2} \le \mu_0 \le \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2} \le \frac{a^2}{4}$$

Set the updating rule

$$\epsilon_k = \min\{\beta a \mu_k, \mu_k^{3/2}\}$$
$$\mu_{k+1} = (2\mu_k^2)^{2/3}\frac{(a+\epsilon_k/\mu_k)^{2/3}}{(a-\epsilon_k/\mu_k)^{4/3}}$$

Then $\mu_{k+1} \le (1-\delta)\mu_k$. *Moreover, for any* $\varepsilon_{\rm dist} > 0$, *running Algorithm 6 after* $K(\mu_0, a, \delta, \varepsilon_{\rm dist})$ *outer iteration*

$$\|W_{\mu_k,\epsilon_k} - W_{\mathsf{G}}\|_2 \le \varepsilon_{\rm dist} \tag{23}$$

where

$$K(\mu_0,a,\delta,\varepsilon_{\rm dist}) \ge \frac{1}{\ln(1/(1-\delta))}\max\left\{\ln\frac{\mu_0}{\beta^2a^2}, \ln\frac{72\mu_0}{a^2(1-(1/2)^{1/4})}, \ln(\frac{3(4-\delta)\mu_0}{{\varepsilon_{\rm dist}}^2}), \frac{1}{2}\ln(\frac{46656\mu_0^2}{a^2{\varepsilon_{\rm dist}}^2}), \frac{1}{3}\ln(\frac{46656\mu_0^3}{a^4{\varepsilon_{\rm dist}}^2})\right\}$$

The total gradient descent steps are

$$\begin{aligned}
&\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\rm dist})} \frac{2(\mu_k(a^2+1)+3a^2)(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2} \\
\le& 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6a^6} + \left(\max\{\frac{3(4-\delta)}{{\varepsilon_{\rm dist}}^2}, \frac{216}{a\varepsilon_{\rm dist}}, \left(\frac{216}{a\varepsilon_{\rm dist}}\right)^{2/3}, \frac{1}{\beta^2a^2}, \frac{72}{(1-(1/2)^{1/4})a^2}\}\right)^3\right) g_{\mu_0}(W_{\mu_0}^{\epsilon_0}) \\
\lesssim& O\left(\mu_0a^2+a^2+\mu_0\right)\left(\frac{1}{\beta^6a^6} + \frac{1}{{\varepsilon_{\rm dist}}^6} + \frac{1}{a^3{\varepsilon_{\rm dist}}^3} + \frac{1}{a^2{\varepsilon_{\rm dist}}^2} + \frac{1}{a^6}\right)
\end{aligned}$$

Proof. Upon substituting gradient flow with gradient descent, it becomes possible to only identify an ϵ-stationary point for $g_\mu(W)$. This modification necessitates specifying the stepsize η for gradient descent, as well as an updating rule for μ. The adjustment procedure used can substantially influence the result of Algorithm 6. In this proof, we will impose limitations on the update scheme μ_k, the stepsize η_k, and the tolerance ϵ_k to ensure their effective operation within Algorithm 6. The approach employed for this proof closely mirrors that of the proof for Theorem 1 albeit with more careful scrutiny. In this proof, we will work out all the requirements for μ, ϵ, η. Subsequently, we will verify that our selection in Theorem 4 conforms to these requirements.

In the proof, we occasionally use μ, ϵ or μ_k, ϵ_k. When we employ μ, ϵ, it signifies that the given inequality or equality holds for any μ, ϵ. Conversely, when we use μ_k, ϵ_k, it indicates we are examining how to set these parameters for distinct iterations.

Establish the Bound $y_{\mu,\epsilon}^{**} \ge \sqrt{\mu}$ First, let us consider $r_\epsilon(\sqrt{\mu};\mu) \le 0$, i.e.

$$r_\epsilon(\sqrt{\mu};\mu) = \frac{a+\epsilon/\mu}{\sqrt{\mu}} - (a^2+1) - \frac{\mu(a-\epsilon/\mu)^2}{4\mu^2} \le 0$$

This is always true when $\mu > 4/a^2$, and we require

$$\epsilon \le 2\mu^{3/2} + a\mu - 2\sqrt{2a\mu^{5/2} - \mu^3a^2} \quad \text{when } \mu \le \frac{4}{a^2}$$

Now we name it condition 1.

Condition 1.

$$\epsilon \le 2\mu^{3/2} + a\mu - 2\sqrt{2a\mu^{5/2} - \mu^3a^2} \quad \textit{when } \mu \le \frac{4}{a^2}$$

Under the assumption that Condition 1 is satisfied. Since $r_\epsilon(y;\mu)$ is increasing function with interval $y \in [y_{{\rm lb},\epsilon}, y_{{\rm ub},\epsilon}]$, and we know $y_{{\rm lb},\epsilon} \le \sqrt{\mu} \le y_{{\rm ub},\epsilon}$ and based on Theorem 7(ii), we have $y_{{\rm lb},\epsilon} \le y_{\mu,\epsilon}^{**} \le y_{{\rm ub},\epsilon}$, $r_\epsilon(\sqrt{\mu};\mu) \le r_\epsilon(y_{\mu,\epsilon}^{**};\mu) = 0$. Therefore, $y_{\mu,\epsilon}^{**} \ge \sqrt{\mu}$.

Ensuring the Correct Solution Path via Gradient Descent Following the argument when we prove Theorem 1, we strive to ensure that the gradient descent, when initiated at $(x_{\mu_k,\epsilon_k}, y_{\mu_k,\epsilon_k})$, will converge within the "correct" ϵ_{k+1}-stationary point region (namely, $\|\nabla g_{\mu_{k+1}}(W)\|_2 < \epsilon_{k+1}$) which includes $(x^*_{\mu_{k+1}}, y^*_{\mu_{k+1}})$. For this to occur, we necessitate that:

$$y_{\mu_{k+1},\epsilon_{k+1}} \overset{(1)}{>} y_{\mu_{k+1},\epsilon_{k+1}}{}^{**} \overset{(2)}{>} \sqrt{\mu_{k+1}} \overset{(3)}{\geq} (2\mu_k^2)^{1/3}\frac{(a+\epsilon_k/\mu_k)^{1/3}}{(a-\epsilon_k/\mu_k)^{2/3}} \overset{(4)}{>} y_{\mu_k,\epsilon_k}{}^{*} \overset{(5)}{>} y_{\mu_k,\epsilon_k} \tag{24}$$

Here (1), (5) are due to Corollary 3; (2) comes from the boundary we established earlier; (3) is based on the constraints we have placed on μ_k and μ_{k+1}, which we will present as Condition 2 subsequently; (4) is from the Theorem 7(ii) and relationship $y^*_{\mu_k,\epsilon_k} < y_{\mathrm{lb},\mu_k,\epsilon_k}$. Also, from the Lemma 9, $\max_{\mu\leq\tau} x^{**}_{\mu,\epsilon} \leq \min_{\mu>0} x^*_{\mu,\epsilon}$. Hence, by invoking Lemma 6, we can affirm that our gradient descent consistently traces the correct stationary point. Now we state condition to make it happen,

Condition 2.

$$(1-\delta)\mu_k \geq \mu_{k+1} \geq (2\mu_k^2)^{2/3}\frac{(a+\epsilon_k/\mu_k)^{2/3}}{(a-\epsilon_k/\mu_k)^{4/3}}$$

In this context, our requirement extends beyond merely ensuring that μ_k decreases. We further stipulate that it should decrease by a factor of $1-\delta$. Next, we impose another important constraint

Condition 3.

$$\epsilon_k \leq \mu_k^{3/2}$$

Updating Rules Now we are ready to check our updating rules satisfy the conditions above

$$\begin{aligned}\epsilon_k =& \min\{\beta a\mu_k, \mu_k^{3/2}\}\\ \mu_{k+1} =& (2\mu_k^2)^{2/3}\frac{(a+\epsilon_k/\mu_k)^{2/3}}{(a-\epsilon_k/\mu_k)^{4/3}}\end{aligned}$$

Check for Conditions First, we check the condition 2, condition 2 requires

$$(1-\delta)\mu_k \geq (2\mu_k^2)^{2/3}\frac{(a+\epsilon_k/\mu_k)^{2/3}}{(a-\epsilon_k/\mu_k)^{4/3}} \Rightarrow \mu_k\frac{(a+\epsilon_k/\mu_k)^2}{(a-\epsilon_k/\mu_k)^4} \leq \frac{(1-\delta)^3}{4}$$

Note that $\epsilon_k \leq \beta a\mu_k < a\mu_k$

$$\mu_k\frac{(a+\epsilon_k/\mu_k)^2}{(a-\epsilon_k/\mu_k)^4} \leq \mu_k\frac{(1+\beta)^2}{(1-\beta)^4}\frac{1}{a^2}$$

Therefore, once the following inequality is true, Condition 2 is satisfied.

$$\mu_k\frac{(1+\beta)^2}{(1-\beta)^4}\frac{1}{a^2} \leq \frac{(1-\delta)^3}{4} \Rightarrow \mu_k \leq \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$$

Because $\mu_k \leq \mu_0 \leq \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$ from the condition we impose for μ_0. Consequently, Condition 2 is satisfied under our choice of ϵ_k.

Now we focus on the Condition 1. Because $\epsilon_k \leq a\beta\mu_k$, if we can ensure $a\beta\mu_k \leq 2\mu_k^{3/2} + a\mu_k - 2\sqrt{2a\mu_k^{5/2} - \mu_k^3a^2}$ holds, then we can show Condition 1 is always satisfied.

$$\begin{aligned}a\beta\mu_k \leq& 2\mu_k^{3/2} + a\mu_k - 2\sqrt{2a\mu_k^{5/2} - \mu_k^3a^2}\\ 2\sqrt{2a\mu_k^{5/2} - \mu_k^3a^2} \leq& 2\mu_k^{3/2} + (1-\beta)a\mu_k\\ 4(2a\mu_k^{5/2} - \mu_k^3a^2) \leq& 4\mu_k^3 + (1-\beta)^2a^2\mu_k^2 + 4(1-\beta)a\mu_k^{5/2}\\ 0 \leq& 4(a^2+1)\mu_k^3 + (1-\beta)^2a^2\mu_k^2 - 4(1+\beta)a\mu_k^{5/2}\\ 0 \leq& 4(a^2+1)\mu_k - 4(1+\beta)a\mu_k^{1/2} + (1-\beta)^2a^2 \qquad \text{when} \quad 0 \leq \mu_k \leq 4/a^2\\ 0 \leq& \mu_k - \frac{(1+\beta)a}{(a^2+1)}\mu_k^{1/2} + \frac{(1-\beta)^2a^2}{4(a^2+1)}\end{aligned}$$

We also notice that

$$\frac{(1+\beta)^2a^2}{(a^2+1)^2} - 4\frac{(1-\beta)^2a^2}{4(a^2+1)} \leq 0 \Leftrightarrow \left(\frac{1+\beta}{1-\beta}\right)^2 \leq a^2+1$$

Because $\left(\frac{1+\beta}{1-\beta}\right)^2 \leq (1-\delta)(a^2+1)$, the inequality above always holds and this inequality implies that for any $\mu_k \geq 0$

$$0 \leq \mu_k - \frac{(1+\beta)a}{(a^2+1)}\mu_k^{1/2} + \frac{(1-\beta)^2a^2}{4(a^2+1)}$$

Therefore, Condition 2 holds. Condition 3 also holds because of the choice of ϵ_k.

Bound the Distance Let $c = 72/a^2$, and assume that μ satisfies the following

$$\mu \leq \min\{\frac{1}{c}\left(1-(1/2)^{1/4}\right), \beta^2a^2\} \tag{25}$$

Note that when μ satisfies (25), then $\mu^{3/2} \leq \beta a\mu$, so $\epsilon = \mu^{3/2}$.

$$\mu \leq \frac{1}{c}\left(1-(1/2)^{1/4}\right) = \frac{a^2}{72}\left(1-(1/2)^{1/4}\right) \leq \frac{a^2}{4}$$

$$\epsilon/\mu = \sqrt{\mu} \leq \frac{a}{2} \tag{26}$$

Then

$$\begin{aligned}
t_\epsilon((a-\epsilon/\mu)(1-c\mu);\mu) =& \frac{1}{1-c\mu} - 1 - \frac{\mu(a+\epsilon/\mu)^2}{(\mu(a^2+1)+(a-\epsilon/\mu)^2(1-c\mu)^2)^2} \\
=& \frac{c\mu}{1-c\mu} - \frac{\mu(a+\epsilon/\mu)^2}{(\mu(a^2+1)+(a-\epsilon/\mu)^2(1-c\mu)^2)^2} \\
\geq& c\mu - \mu\frac{(a+\epsilon/\mu)^2}{(a-\epsilon/\mu)^4(1-c\mu)^4} \\
\geq& c\mu - \mu\frac{(a+a/2)^2}{(a-a/2)^4(1-c\mu)^4} \\
=& \mu\left(c - \frac{36}{a^2(1-c\mu)^4}\right) \\
=& \mu\left(\frac{72}{a^2} - \frac{36}{a^2(1-c\mu)^4}\right) > 0
\end{aligned}$$

Then we know $(a-\epsilon/\mu)(1-c\mu) < x^*_{\mu,\epsilon}$. Now we can bound the distance $\|W_{\mu_k,\epsilon_k} - W_\mathsf{G}\|$, it is important to note that

$$\begin{aligned}
\|W_{\mu_k,\epsilon_k} - W_\mathsf{G}\| =& \sqrt{(x_{\mu_k,\epsilon_k}-a)^2 + (y_{\mu_k,\epsilon_k})^2} \\
\leq& \max\left\{\sqrt{(x^*_{\mu_k,\epsilon_k}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2}, \sqrt{(x^*_{\mu_k,\epsilon_{k_-}}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2}\right\}
\end{aligned}$$

We use the fact that $x^*_{\mu_k,\epsilon_k} < x_{\mu_k,\epsilon_k} < a$, $x_{\mu_k,\epsilon_k} < x^*_{\mu_k,\epsilon_{k_-}}$ and $y_{\mu_k,\epsilon_k} < y^*_{\mu_k,\epsilon_k}$. Next, we can separately establish bounds for these two terms. Due to (24), $y^*_{\mu_k,\epsilon_k} < (2\mu_k^2)^{1/3}\frac{(a+\epsilon_k/\mu_k)^{1/3}}{(a-\epsilon_k/\mu_k)^{2/3}} = \sqrt{\mu_{k+1}}$ and $(a-\epsilon_k/\mu_k)(1-c\mu_k) < x^*_{\mu_k,\epsilon_k}$

$$\sqrt{(x^*_{\mu_k,\epsilon_k}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2} \leq \sqrt{\mu_{k+1} + (a-(a-\epsilon_k/\mu_k)(1-c\mu_k))^2}$$

Given that if $x^*_{\mu_k,\epsilon_{k_-}} \leq a$, then $\sqrt{(x^*_{\mu_k,\epsilon_k}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2} \geq \sqrt{(x^*_{\mu_k,\epsilon_{k_-}}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2}$. Therefore, if $x^*_{\mu_k,\epsilon_{k_-}} \geq a$, we can use the fact that $x^*_{\mu_k,\epsilon_{k_-}} \leq a + \frac{\epsilon_k}{\mu_k}$. In this case,

$$\sqrt{(x^*_{\mu_k,\epsilon_{k_-}}-a)^2 + (y^*_{\mu_k,\epsilon_k})^2} \leq \sqrt{\mu_{k+1} + (\epsilon_k/\mu_k)^2} = \sqrt{\mu_{k+1}+\mu_k} \leq \sqrt{(2-\delta)\mu_k}$$

As a result, we have

$$\|W_{\mu_k,\epsilon_k} - W_{\mathsf{G}}\| \leq \max\{\sqrt{\mu_{k+1} + (a - (a - \epsilon_k/\mu_k)(1 - c\mu_k))^2}, \sqrt{(2-\delta)\mu_k}\}$$

$$\begin{aligned}\mu_{k+1} + (a - (a - \epsilon_k/\mu_k)(1 - c\mu_k))^2 \leq& (1-\delta)\mu_k + (ac\mu_k + \sqrt{\mu_k} - c\mu_k^{3/2})^2 \\ \leq& (1-\delta)\mu_k + 3(a^2c^2\mu_k^2 + \mu_k + c^2\mu_k^3) \\ =& (4-\delta)\mu_k + 3a^2c^2\mu_k^2 + 3c^2\mu_k^3\end{aligned}$$

$$\begin{aligned}\|W_{\mu_k,\epsilon_k} - W_{\mathsf{G}}\| \leq& \max\{\sqrt{\mu_{k+1} + (a - (a - \epsilon_k/\mu_k)(1 - c\mu_k))^2}, \sqrt{(2-\delta)\mu_k}\} \\ \leq& \max\{\sqrt{(4-\delta)\mu_k + 3a^2c^2\mu_k^2 + 3c^2\mu_k^3}, \sqrt{(2-\delta)\mu_k}\} \\ =& \sqrt{(4-\delta)\mu_k + 3a^2c^2\mu_k^2 + 3c^2\mu_k^3}\end{aligned}$$

Just let

$$(4-\delta)\mu_k \leq (4-\delta)(1-\delta)^k\mu_0 \leq \frac{{\varepsilon_{\text{dist}}}^2}{3} \Rightarrow k \geq \frac{\ln(3(4-\delta)\mu_0/{\varepsilon_{\text{dist}}}^2)}{\ln(1/(1-\delta))} \tag{27}$$

$$3a^2c^2\mu_k^2 \leq 3a^2c^2(1-\delta)^{2k}\mu_0^2 \leq \frac{{\varepsilon_{\text{dist}}}^2}{3} \Rightarrow k \geq \frac{\ln(46656\mu_0^2/(a^2{\varepsilon_{\text{dist}}}^2))}{2\ln(1/(1-\delta))} \tag{28}$$

$$3c^2\mu_k^3 \leq 3c^2(1-\delta)^{3k}\mu_0^3 \leq \frac{{\varepsilon_{\text{dist}}}^2}{3} \Rightarrow k \geq \frac{\ln(46656\mu_0^3/(a^4{\varepsilon_{\text{dist}}}^2))}{3\ln(1/(1-\delta))} \tag{29}$$

We use the fact that $\mu_k \leq (1-\delta)^k\mu_0$. In order to satisfy (25).

$$\mu_k \leq \mu_0(1-\delta)^k \leq \frac{a^2}{72}(1-(1/2)^{1/4}) \Rightarrow k \geq \frac{\ln\frac{72\mu_0}{a^2(1-(1/2)^{1/4})}}{\ln\frac{1}{1-\delta}} \tag{30}$$

$$\mu_k \leq \mu_0(1-\delta)^k \leq \beta^2a^2 \Rightarrow k \geq \frac{\ln\left(\mu_0/(\beta^2a^2)\right)}{\ln\frac{1}{1-\delta}} \tag{31}$$

Consequently, running Algorithm 6 after $K(\mu_0, a, \delta, \varepsilon_{\text{dist}})$ outer iteration

$$\|W_{\mu_k,\epsilon_k} - W_{\mathsf{G}}\|_2 \leq \varepsilon_{\text{dist}}$$

where

$$K(\mu_0, a, \delta, \varepsilon_{\text{dist}}) \geq \frac{1}{\ln(1/(1-\delta))}\max\left\{\ln\frac{\mu_0}{\beta^2a^2}, \ln\frac{72\mu_0}{a^2(1-(1/2)^{1/4})}, \ln(\frac{3(4-\delta)\mu_0}{\varepsilon^2}), \frac{1}{2}\ln(\frac{46656\mu_0^2}{a^2\varepsilon^2}), \frac{1}{3}\ln(\frac{46656\mu_0^3}{a^4\varepsilon^2})\right\}$$

By Lemma 6, k iteration of Algorithm 6 need the following step of gradient descent

$$\frac{2(\mu_k(a^2+1)+3a^2)(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2}$$

Let $\widehat{K}(\mu_0, a, \delta, \varepsilon_{\text{dist}})$ satisfy $\mu_{\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \leq \beta^2 a^2 < \mu_{\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})-1}$. Hence, the total number of gradient steps required by Algorithm 6 can be expressed as follows:

$$
\begin{aligned}
&\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{2(\mu_k(a^2+1)+3a^2)(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2} \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\sum_{k=0}^{\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})-1} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2} + \sum_{k=\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2}\right) \\
= & 2(\mu_0(a^2+1)+3a^2)\left(\sum_{k=0}^{\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})-1} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\beta^2 a^2 \mu_k^2} + \sum_{k=\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\mu_k^3}\right) \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\sum_{k=0}^{\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})-1} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\beta^6 a^6} + \sum_{k=\widehat{K}(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\mu_k^3}\right) \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\beta^6 a^6} + \sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^3}\right) \\
= & 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6 a^6} + \frac{1}{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^3}\right)\left(\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \left((g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))\right)\right) \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6 a^6} + \frac{1}{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^3}\right)\left(\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \left((g_{\mu_k}(W_{\mu_k}^{\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1}}^{\epsilon_{k+1}}))\right)\right) \\
= & 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6 a^6} + \frac{1}{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^3}\right)\left((g_{\mu_0}(W_{\mu_0,\epsilon_0}) - g_{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})+1}}(W_{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})+1}}^{\epsilon_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})+1}})\right) \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6 a^6} + \frac{1}{\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})}^3}\right) g_{\mu_0}(W_{\mu_0,\epsilon_0})
\end{aligned}
$$

Note from (27) and (30), the following should holds

$$
\mu_{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} = \min\{\frac{{\varepsilon_{\text{dist}}}^2}{3(4-\delta)}, \frac{a\varepsilon_{\text{dist}}}{216}, \left(\frac{a\varepsilon_{\text{dist}}}{216}\right)^{2/3}, \beta^2 a^2, \frac{a^2}{72}(1-(1/2)^{1/4})\}
$$

Therefore,

$$
\begin{aligned}
&\sum_{k=0}^{K(\mu_0,a,\delta,\varepsilon_{\text{dist}})} \frac{2(\mu_k(a^2+1)+3a^2)(g_{\mu_{k+1}}(W_{\mu_k,\epsilon_k}) - g_{\mu_{k+1}}(W_{\mu_{k+1},\epsilon_{k+1}}))}{\epsilon_k^2} \\
\leq & 2(\mu_0(a^2+1)+3a^2)\left(\frac{1}{\beta^6 a^6} + \left(\max\{\frac{3(4-\delta)}{{\varepsilon_{\text{dist}}}^2}, \frac{216}{a\varepsilon_{\text{dist}}}, \left(\frac{216}{a\varepsilon_{\text{dist}}}\right)^{2/3}, \frac{1}{\beta^2 a^2}, \frac{72}{(1-(1/2)^{1/4})a^2}\}\right)^3\right) g_{\mu_0}(W_{\mu_0}^{\epsilon_0})
\end{aligned}
$$

□

D Additional Theorems and Lemmas

Theorem 5 (Detailed Property of $r(y;\mu)$)**.** *For $r(y;\mu)$ in (8), then*

(i) *For $\mu > 0$, $\lim_{y\to 0^+} r(y;\mu) = \infty$, $r(\frac{a}{a^2+1},\mu) < 0$*

(ii) *For $\mu > 0$, $r(\sqrt{\mu},\mu) < 0$.*

(iii) *For $\mu > \frac{a^2}{4}$*

$$\frac{dr(y;\mu)}{dy} < 0$$

For $0 < \mu \le \frac{a^2}{4}$

$$\begin{cases} \dfrac{dr(y;\mu)}{dy} > 0 & y_{\text{lb}} < y < y_{\text{ub}} \quad (32a) \\ \dfrac{dr(y;\mu)}{dy} \le 0 & \textit{Otherwise} \quad (32b) \end{cases}$$

where

$$y_{\text{lb}} = \frac{(4\mu)^{1/3}}{2}(a^{1/3} - \sqrt{a^{2/3} - (4\mu)^{1/3}}) \quad y_{\text{ub}} = \frac{(4\mu)^{1/3}}{2}(a^{1/3} + \sqrt{a^{2/3} - (4\mu)^{1/3}})$$

Moreover,

$$y_{\text{lb}} \le \sqrt{\mu} \le y_{\text{ub}}$$

(iv) *For $0 < \mu < \frac{a^2}{4}$, let $p(\mu) = r(y_{\text{ub}},\mu)$, then $p'(\mu) < 0$ and there exist a unique solution to $p(\mu) = 0$, denoted as τ. Additionally, $\tau < \frac{a^2}{4}$.*

(v) *There exists a $\tau > 0$ such that, $\forall \mu > \tau$, the equation $r(y;\mu) = 0$ has only one solution. At $\mu = \tau$, the equation $r(y;\mu) = 0$ has two solutions, and $\forall \mu < \tau$, the equation $r(y;\mu) = 0$ has three solutions. Moreover, $\mu < \frac{a^2}{4}$.*

(vi) *$\forall \mu < \tau$, the equation $r(y;\mu) = 0$ has three solution, i.e. $y_\mu^* < y_\mu^{**} < y_\mu^{***}$.*

$$\frac{dy_\mu^*}{d\mu} > 0 \quad \frac{dy_\mu^{**}}{d\mu} > 0 \quad \frac{dy_\mu^{***}}{d\mu} < 0 \text{ and } \lim_{\mu\to 0} y_\mu^* = 0, \lim_{\mu\to 0} y_\mu^{**} = 0, \lim_{\mu\to 0} y_\mu^{***} = \frac{a}{a^2+1}$$

Moreover,

$$y_\mu^* < y_{\text{lb}} < \sqrt{\mu} < y_\mu^{**} < y_{\text{ub}} < y_\mu^{***}$$

Theorem 6 (Detailed Property of $t(x;\mu)$)**.** *For $t(x;\mu)$ in (9), then*

(i) *For $\mu > 0$, $\lim_{x\to 0^+} t(x;\mu) = \infty$, $t(a,\mu) < 0$*

(ii) *If $\mu < \left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2$ or $\mu > \left(\frac{a(\sqrt{a^2+1}+a)}{2(a^2+1)}\right)^2$, then $t(\sqrt{\mu(a^2+1)},\mu) < 0$.*

(iii) *For $\mu > \frac{a^2}{4(a^2+1)^3}$*

$$\frac{dt(x;\mu)}{dx} < 0$$

For $0 < \mu \le \frac{a^2}{4(a^2+1)^3}$

$$\begin{cases} \dfrac{dt(x;\mu)}{dx} > 0 & x_{\text{lb}} < x < x_{\text{ub}} \quad (33a) \\ \dfrac{dt(x;\mu)}{dx} \le 0 & \textit{Otherwise} \quad (33b) \end{cases}$$

where

$$x_{\text{lb}} = \frac{(4\mu a)^{1/3}(1-\sqrt{1-\frac{(4\mu)^{1/3}(a^2+1)}{a^{2/3}}})}{2} \quad x_{\text{ub}} = \frac{(4\mu a)^{1/3}(1+\sqrt{1-\frac{(4\mu)^{1/3}(a^2+1)}{a^{2/3}}})}{2}$$

Moreover,

$$x_{\text{lb}} \le \sqrt{\mu(a^2+1)} \le x_{\text{ub}}$$

(iv) *For* $0 < \mu < \frac{a^2}{4(a^2+1)^3}$ *and let* $q(\mu) = t(x_{\text{lb}}, \mu)$*, then* $q'(\mu) > 0$ *and there exist a unique solution to* $q(\mu) = 0$*, denoted as* τ *and* $\tau < \frac{a^2}{4(a^2+1)^3} \le \frac{1}{27}$.

(v) *There exists a* $\tau > 0$ *such that,* $\forall \mu > \tau$*, the equation* $t(x;\mu) = 0$ *has only one solution. At* $\mu = \tau$*, the equation* $t(x;\mu) = 0$ *has two solutions, and* $\forall \mu < \tau$*, the equation* $t(x;\mu) = 0$ *has three solutions. Moreover,* $\tau < \frac{a^2}{4(a^2+1)^3} \le \frac{1}{27}$

(vi) $\forall \mu < \tau$*,* $t(x;\mu) = 0$ *has three stationary points, i.e.* $x_\mu^{***} < x_\mu^{**} < x_\mu^*$.

$$\frac{dx_\mu^*}{d\mu} < 0 \quad \frac{dx_\mu^{***}}{d\mu} > 0 \text{ and } \lim_{\mu\to 0} x_\mu^* = a, \lim_{\mu\to 0} x_\mu^{**} = 0, \lim_{\mu\to 0} x_\mu^{***} = 0$$

Besides,

$$\max_{\mu\le\tau} x_\mu^{**} \le \frac{a(\sqrt{a^2+1}-a)}{2\sqrt{a^2+1}} \quad \text{and} \quad \frac{a(\sqrt{a^2+1}+a)}{2\sqrt{a^2+1}} \le \min_{\mu>0} x_\mu^*$$

It also implies that $t(\frac{a(\sqrt{a^2+1}-a)}{2\sqrt{a^2+1}};\mu) \ge 0$ *and* $\max_{\mu\le\mu_0} x_\mu^{**} < \min_{\mu>0} x_\mu^*$

Lemma 7. *Algorithm 1 with input* $f = g_\mu(x,y), \boldsymbol{z}_0 = (x(0), y(0))$ *where* $(x(0), y(0)) \in C_{\mu 3}$ *in (41), then* $\forall t \ge 0$*,* $(x(t), y(t)) \in C_{\mu 3}$*. Moreover,* $\lim_{t\to\infty}(x(t), y(t)) = (x_\mu^*, y_\mu^*)$

Lemma 8. *For any* $(x,y) \in C_{\mu 3}$ *in (41), and* $(x,y) \ne (x_\mu^*, y_\mu^*)$

$$g_\mu(x,y) > g_\mu(x_\mu^*, y_\mu^*)$$

Theorem 7 (Detailed Property of $r_\epsilon(y;\mu)$)**.** *For* $r_\epsilon(y;\mu)$ *in (15), then*

(i) *For* $\mu > 0, \epsilon > 0$*,* $\lim_{y\to 0^+} r_\epsilon(y;\mu) = \infty$*,* $y(\frac{a}{a^2+1}, \mu) < 0$

(ii) *For* $\mu > \frac{(a-\epsilon/\mu)^4}{4(a+\epsilon/\mu)^2}$*, then* $\frac{dr_\epsilon(y;\mu)}{dy} < 0$*. For* $0 < \mu \le \frac{(a-\epsilon/\mu)^4}{4(a+\epsilon/\mu)^2}$

$$\begin{cases} \dfrac{dr_\epsilon(y;\mu)}{dy} > 0 & y_{\text{lb},\mu,\epsilon} < y < y_{\text{ub},\mu,\epsilon} \quad (34a) \\ \dfrac{dr_\epsilon(y;\mu)}{dy} \le 0 & \textit{Otherwise} \quad (34b) \end{cases}$$

where

$$y_{\text{lb},\mu,\epsilon} = \frac{(4\mu)^{1/3}}{2}\left(\left(\frac{(a-\epsilon/\mu)^2}{a-\epsilon/\mu}\right)^{1/3} - \sqrt{\left(\frac{(a-\epsilon/\mu)^2}{a-\epsilon/\mu}\right)^{2/3} - (4\mu)^{1/3}}\right)$$

$$y_{\text{ub},\mu,\epsilon} = \frac{(4\mu)^{1/3}}{2}\left(\left(\frac{(a-\epsilon/\mu)^2}{a-\epsilon/\mu}\right)^{1/3} + \sqrt{\left(\frac{(a-\epsilon/\mu)^2}{a-\epsilon/\mu}\right)^{2/3} - (4\mu)^{1/3}}\right)$$

Also,

$$y_{\text{lb},\mu,\epsilon} \le (2\mu^2)^{1/3}\frac{(a+\epsilon/\mu)^{1/3}}{(a-\epsilon/\mu)^{2/3}}$$

$$y_{\text{lb},\mu,\epsilon} \le \sqrt{\mu} \le y_{\text{ub},\mu,\epsilon}$$

Theorem 8 (Detailed Property of $r_\beta(y;\mu)$)**.** *For* $r_\beta(y;\mu)$ *in (19), then*

(i) *For* $\mu > 0, \epsilon > 0$, $\lim_{y\to 0^+} r_\beta(y;\mu) = \infty$

(ii) *For* $\mu > \frac{a^2(1-\beta)^4}{4(1+\beta)^2}$, *then* $\frac{dr_\beta(y;\mu)}{dy} < 0$. *For* $0 < \mu \le \frac{a^2(1-\beta)^4}{4(1+\beta)^2}$

$$\begin{cases} \dfrac{dr_\beta(y;\mu)}{dy} > 0 & y_{\mathrm{lb},\mu,\beta} < y < y_{\mathrm{ub},\mu,\beta} \quad (35a) \\ \dfrac{dr_\beta(y;\mu)}{dy} \le 0 & \textit{Otherwise} \quad (35b) \end{cases}$$

where

$$y_{\mathrm{lb},\mu,\beta} = \frac{(4\mu)^{1/3}}{2}\left(\frac{a(1-\beta)^2}{1+\beta}\right)^{1/3}\left(1-\sqrt{1-\frac{(4\mu)^{1/3}}{a^{2/3}}\left(\frac{1+\beta}{(1-\beta)^2}\right)^{2/3}}\right)$$

$$y_{\mathrm{ub},\mu,\beta} = \frac{(4\mu)^{1/3}}{2}\left(\frac{a(1-\beta)^2}{1+\beta}\right)^{1/3}\left(1+\sqrt{1-\frac{(4\mu)^{1/3}}{a^{2/3}}\left(\frac{1+\beta}{(1-\beta)^2}\right)^{2/3}}\right)$$

Also,

$$y_{\mathrm{lb},\mu,\beta} \le \frac{(4\mu)^{2/3}}{2a^{1/3}}\frac{(1+\beta)^{1/3}}{(1-\beta)^{2/3}}$$

$$y_{\mathrm{lb},\mu,\beta} \le \sqrt{\mu} \le y_{\mathrm{ub},\mu,\beta}$$

Theorem 9 (Detailed Property of $t_\beta(x;\mu)$)**.** *For* $t_\beta(x;\mu)$ *in* (20)*, then*

(i) *For* $\mu > 0$, $\lim_{x\to 0^+} t_\beta(x;\mu) = \infty$, $t_\beta(a;\mu) < 0$

(ii) *For* $\mu > \frac{a^2}{4(a^2+1)^3}\frac{(\beta+1)^4}{(\beta-1)^2}$

$$\frac{dt_\beta(x;\mu)}{dx} < 0$$

For $0 < \mu \le \frac{a^2}{4(a^2+1)^3}\frac{(\beta+1)^4}{(\beta-1)^2}$

$$\begin{cases} \dfrac{dt_\beta(x;\mu)}{dx} > 0 & x_{\mathrm{lb},\mu,\beta} < x < x_{\mathrm{ub},\mu,\beta} \quad (36a) \\ \dfrac{dt_\beta(x;\mu)}{dx} \le 0 & \textit{Otherwise} \quad (36b) \end{cases}$$

where

$$x_{\mathrm{lb},\mu,\beta} = \frac{1}{2}\left(\frac{4a\mu(1+\beta)^2}{1-\beta}\right)^{1/3}\left(1-\sqrt{1-\frac{(4\mu)^{1/3}(a^2+1)}{a^{2/3}}\left(\frac{1-\beta}{(1+\beta)^2}\right)^{2/3}}\right)$$

$$x_{\mathrm{ub},\mu,\beta} = \frac{1}{2}\left(\frac{4a\mu(1+\beta)^2}{1-\beta}\right)^{1/3}\left(1+\sqrt{1-\frac{(4\mu)^{1/3}(a^2+1)}{a^{2/3}}\left(\frac{1-\beta}{(1+\beta)^2}\right)^{2/3}}\right)$$

(iii) *If* $0 < \beta < \frac{\sqrt{(a^2+1)}-1}{\sqrt{(a^2+1)}+1}$, *then there exists a* $\tau_\beta > 0$ *such that,* $\forall \mu > \tau_\beta$, *the equation* $r_\beta(x;\mu) = 0$ *has only one solution. At* $\mu = \tau_\beta$, *the equation* $r_\beta(x;\mu) = 0$ *has two solutions, and* $\forall \mu < \tau_\beta$, *the equation* $r_\beta(x;\mu) = 0$ *has three solutions. Moreover,* $\mu < \frac{a^2}{4(a^2+1)^3}\frac{(\beta+1)^4}{(\beta-1)^2}$.

(iv) *If* $0 < \beta < \frac{\sqrt{(a^2+1)}-1}{\sqrt{(a^2+1)}+1}$, *then* $\forall \mu < \tau_\beta$, $t_\beta(x;\mu) = 0$ *has three stationary points, i.e.* $x^{***}_{\mu,\beta} < x^{**}_{\mu,\beta} < x^{*}_{\mu,\beta}$. *Besides,*

$$\max_{\mu\le\tau_\beta} x^{**}_{\mu,\beta} \le \frac{a((1-\beta)\sqrt{a^2+1} - \sqrt{(1-\beta)^2(a^2+1)-(\beta+1)^2})}{2\sqrt{a^2+1}}$$

$$\frac{a((1-\beta)\sqrt{a^2+1} + \sqrt{(1-\beta)^2(a^2+1)-(\beta+1)^2})}{2\sqrt{a^2+1}} \le \min_{\mu>0} x^{*}_{\mu,\beta}$$

It implies that

$$\max_{\mu \leq \tau_\beta} x^{**}_{\mu,\beta} < \min_{\mu > 0} x^{*}_{\mu,\beta}$$

Lemma 9. *Under the same setting as Corollary 3*

$$\max_{\mu \leq \tau} x^{**}_{\mu,\epsilon} < \min_{\mu > 0} x^{*}_{\mu,\epsilon}$$

E Technical Proofs

E.1 Proof of Theorem 3

Proof. For the sake of completeness, we have included the proof here. Please note that this proof can also be found in [33].

Proof. We use the fact that f is L-smooth function if and only if for any $W, Y \in \text{dom}(f)$

$$f(W) \leq f(Y) + \langle \nabla f(Y), Y - W \rangle + \frac{L}{2}\|Y - W\|_2^2$$

Let $W = W^{t+1}$ and $Y = W^t$, then using the updating rule $W^{t+1} = W^t - \frac{1}{L}\nabla f(W^t)$

$$\begin{aligned} f(W^{t+1}) \leq & f(W^t) + \langle \nabla f(W^t), W^{t+1} - W^t \rangle + \frac{L}{2}\|W^{t+1} - W^t\|_2^2 \\ = & f(W^t) - \frac{1}{L}\|\nabla f(W^t)\|_2^2 + \frac{1}{2L}\|\nabla f(W^t)\|_2^2 \\ = & f(W^t) - \frac{1}{2L}\|\nabla f(W^t)\|_2^2 \end{aligned}$$

Therefore,

$$\min_{0 \leq t \leq n-1} \|\nabla f(W^t)\|_2^2 \leq \frac{1}{n}\sum_{t=0}^{n-1} \|\nabla f(W^t)\|_2^2 \leq \frac{2L(f(W^0) - f(W^n))}{n} \leq \frac{2L(f(W^0) - f(W^*))}{n}$$

$$\min_{0 \leq t \leq n-1} \|\nabla f(W^t)\|_2^2 \leq \frac{2L(f(W^0) - f(W^*))}{n} \leq \epsilon^2 \Rightarrow n \geq \frac{2L(f(W^0) - f(W^*))}{\epsilon^2}$$

□

□

E.2 Proof of Theorem 5

Proof. (i) For any $\mu > 0$,

$$\begin{aligned} \lim_{y \to 0^+} r(y; \mu) &= \lim_{y \to 0^+} \frac{a}{y} - \frac{a^2}{\mu} - (a^2 + 1) = \infty \\ r(\frac{a}{a^2+1}) &= -\frac{\mu a^2}{(\frac{a}{a^2+1})^2 + \mu} < 0. \end{aligned}$$

(ii)

$$\begin{aligned} r(\sqrt{\mu}, \mu) = & \frac{a}{\sqrt{\mu}} - \frac{a^2}{4\mu} - (a^2 + 1) \\ = & -\frac{a^2}{4}(\frac{1}{\sqrt{\mu}} - \frac{2}{a})^2 - a^2 < 0 \end{aligned}$$

(iii)

$$\begin{aligned}\frac{dr(y;\mu)}{dy} =& -\frac{a}{y^2}+\frac{4a^2\mu y}{(y^2+\mu)^3}\\ =&\frac{4a^2\mu y^3-a(y^2+\mu)^3}{y^2(y^2+\mu)^3}\\ =&\frac{a((4a\mu)^{2/3}y^2+(4a\mu)^{1/3}y(y^2+\mu)+(y^2+\mu)^2)((4a\mu)^{1/3}y-y^2-\mu)}{y^2(y^2+\mu)^3}\end{aligned}$$

For $\mu \geq \frac{a^2}{4}$, $((4a\mu)^{1/3}y-y^2-\mu)<0 \Leftrightarrow \frac{dr(y;\mu)}{dy}<0$.

For $\mu < \frac{a^2}{4}$, $y_{\text{lb}} < y < y_{\text{ub}}, ((4a\mu)^{1/3}y-y^2-\mu)>0 \Leftrightarrow \frac{dr(y;\mu)}{dy}>0$. For $\mu<\frac{a^2}{4}$, $y<y_{\text{lb}}$ or $y_{\text{ub}}<y, ((4a\mu)^{1/3}y-y^2-\mu)\leq 0 \Leftrightarrow \frac{dr(y;\mu)}{dy}\leq 0$.

Note that

$$\frac{dr(y;\mu)}{d\mu}=0 \Leftrightarrow ((4a\mu)^{1/3}y-y^2-\mu)=0 \Leftrightarrow (4a\mu)^{1/3}=y+\frac{\mu}{y}$$

The intersection between line $(4a\mu)^{1/3}$ and function $y+\frac{\mu}{y}$ are exactly y_{lb} and y_{ub}, and $y_{\text{lb}}<\sqrt{\mu}<y_{\text{ub}}$.

(iv) Note that for $0<\mu<\frac{a^2}{4}$,

$$\frac{\partial r}{\partial \mu}=-a^2\frac{y^2-\mu}{(\mu+y^2)^3} \quad \text{and} \quad y_{\text{lb}}<\sqrt{\mu}<y_{\text{ub}}$$

then $\frac{\partial r}{\partial \mu}\Big|_{y=y_{\text{ub}}}<0$. Let $p(\mu)=r(y_{\text{ub}},\mu)$, because $\frac{\partial r}{\partial y}|_{y=y_{\text{ub}}}=0$, then

$$\frac{dp(\mu)}{d\mu}=\frac{dr(y_{\text{ub}},\mu)}{d\mu}=\frac{\partial r}{\partial y}\Big|_{y=y_{\text{ub}}}\frac{dy_{\text{ub}}}{d\mu}+\frac{\partial r}{\partial \mu}\Big|_{y=y_{\text{ub}}}=\frac{\partial r}{\partial \mu}\Big|_{y=y_{\text{ub}}}<0$$

Also note that when $\mu=\frac{a^2}{4}$, $y_{\text{ub}}=\sqrt{\mu}$, $p(\mu)=r(y_{\text{ub}},\mu)=r(\sqrt{\mu},\mu)<0$, and also if $\mu<\frac{a^2}{4}$, then

$$y_{\text{ub}}<\frac{(4\mu)^{1/3}}{2}2a^{1/3}=(4\mu a)^{1/3}$$

Thus,

$$\begin{aligned}r((4\mu a)^{1/3},\mu)=&\frac{a}{(4\mu a)^{1/3}}-\frac{\mu a^2}{((4\mu a)^{2/3}+\mu)^2}-(a^2+1)\\ =&\frac{a}{(4\mu a)^{1/3}}-\frac{a^2}{(\mu)^{1/3}((4a)^{2/3}+\mu^{1/3})^2}-(a^2+1)\\ >&\frac{1}{\mu^{1/3}}(\frac{a}{(4a)^{1/3}}-\frac{a^2}{(4a)^{4/3}})-(a^2+1)\end{aligned}$$

Because $\frac{a}{(4a)^{1/3}}>\frac{a^2}{(4a)^{4/3}}$, it is easy to see when $\mu\to 0$, $r((4\mu a)^{1/3},\mu)\to\infty$. We know $r(y_{\text{ub}},\mu)>r((4\mu a)^{1/3},\mu)\to\infty$ as $\mu\to 0$ because of the monotonicity of $r(y;\mu)$ in Theorem 5(iii). Combining all of these, i.e.

$$\frac{dp(\mu)}{d\mu}<0, \quad \lim_{\mu\to 0^+}p(\mu)=\infty, \quad p(\frac{a^2}{4})<0$$

There exists a $\tau<\frac{a^2}{4}$ such that $p(\tau)=0$

(v) From Theorem 5(iv), for $\mu>\tau$, then $p(\mu)=r(y_{\text{ub}},\mu)>0$, and for $\mu=\tau$, then $p(\mu)=r(y_{\text{ub}},\mu)=0$. For $\mu<\tau$, then $p(\mu)=r(y_{\text{ub}},\mu)<0$, combining Theorem 5(i),5(iii), we get the conclusions.

(vi) By Theorem 5(v), $\forall \mu < \tau$, there exists three stationary points such that $0 < y_\mu^* < y_{\text{lb}} < \sqrt{\mu} < y_\mu^{**} < y_{\text{ub}} < y_\mu^{***}$. Because $\frac{dr(y;\mu)}{dy}\Big|_{y=y_{\text{lb}}} = \frac{dr(y;\mu)}{dy}\Big|_{y=y_{\text{ub}}} = 0$, then

$$\frac{dr(y;\mu)}{dy}\Big|_{y=y_\mu^*} \neq 0, \quad \frac{dr(y;\mu)}{dy}\Big|_{y=y_\mu^{**}} \neq 0, \quad \frac{dr(y;\mu)}{dy}\Big|_{y=y_\mu^{***}} \neq 0$$

By implicit function theorem [14], for solution to equation $r(y;\mu) = 0$, there exists a unique continuously differentiable function such that $y = y(\mu)$ and satisfies $r(y(\mu),\mu) = 0$. Therefore,

$$\frac{\partial r}{\partial \mu} = -a^2\frac{y^2-\mu}{(\mu+y^2)^3}, \quad \frac{\partial r}{\partial y} = -\frac{a}{y^2} + \frac{4a^2\mu y}{(y^2+\mu)^3}, \quad \frac{dy(\mu)}{d\mu} = -\frac{\partial r/\partial \mu}{\partial r/\partial y}$$

Therefore by Theorem 5(iii),

$$\frac{dy}{d\mu}\Big|_{y=y_\mu^*} > 0 \quad \frac{dy}{d\mu}\Big|_{y=y_\mu^{**}} > 0 \quad \frac{dy}{d\mu}\Big|_{y=y_\mu^{***}} < 0$$

Because $\lim_{\mu\to 0^+} y_{\text{lb}} = \lim_{\mu\to 0^+} y_{\text{ub}} = 0$, then $\lim_{\mu\to 0^+} y_\mu^* = \lim_{\mu\to 0^+} y_\mu^{**} = 0$. Let us consider $r(\frac{a}{a^2+1}(1-c\mu),\mu)$ where $c = 32\frac{(a^2+1)^3}{a^2}$ and $\mu < \frac{1}{2c}$

$$\begin{aligned}
&r(\frac{a}{a^2+1}(1-c\mu),\mu)\\
=&\frac{a}{\frac{a}{a^2+1}(1-c\mu)} - \frac{\mu a^2}{(\frac{a^2}{(a^2+1)^2}(1-c\mu)^2+\mu)^2} - (a^2+1)\\
=&(a^2+1)(\frac{c\mu}{1-c\mu}) - \frac{\mu a^2}{(\frac{a^2}{(a^2+1)^2}(1-c\mu)^2+\mu)^2}\\
\geq& c(a^2+1)\mu - \frac{\mu a^2}{(\frac{a^2}{(a^2+1)^2}(1-c\mu)^2)^2}\\
=&c(a^2+1)\mu - \frac{16(a^2+1)^4}{a^2}\mu\\
=&\frac{16(a^2+1)^4}{a^2}\mu > 0
\end{aligned}$$

By Theorem 5(iii), then $\frac{a}{a^2+1}(1-c\mu) < y_\mu^{***}$, then

$$\frac{a}{a^2+1} = \lim_{\mu\to 0^+}\frac{a}{a^2+1}(1-c\mu),\mu) \leq \lim_{\mu\to 0^+} y_\mu^{***} \leq \frac{a}{a^2+1}$$

Consequently,

$$\lim_{\mu\to 0^+} y_\mu^{***} = \frac{a}{a^2+1}$$

□

E.3 Proof of Theorem 6

Proof. (i) For $\mu > 0$,

$$\lim_{x\to 0^+} t(x;\mu) = \lim_{x\to 0^+}\frac{a}{x} - \frac{a^2}{\mu(a^2+1)^2} - 1 = \infty$$

$$t(a,\mu) = -\frac{\mu a^2}{(\mu(a^2+1)+a^2)^2} < 0$$

(ii)

$$t(\sqrt{\mu(a^2+1)},\mu) = \frac{a}{\sqrt{a^2+1}}\frac{1}{\sqrt{\mu}} - \frac{a^2}{4\mu(a^2+1)^2} - 1$$

If $t(\sqrt{\mu(a^2+1)},\mu)=0$, then

$$\frac{1}{\sqrt{\mu}}=2\frac{(a^2+1)^{3/2}}{a}\pm 2(a^2+1)\Rightarrow \mu=\left(\frac{a(\sqrt{a^2+1}\mp a)}{2(a^2+1)}\right)^2$$

so when $\mu<\left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2$ or $\mu>\left(\frac{a(\sqrt{a^2+1}+a)}{2(a^2+1)}\right)^2$, then $t(\sqrt{\mu(a^2+1)},\mu)<0$

(iii)

$$\begin{aligned}
&\frac{dt(x,\mu)}{dx}\\
=&-\frac{a}{x^2}+\frac{4\mu a^2 x}{(\mu(a^2+1)+x^2)^3}\\
=&\frac{4\mu a^2x^3-a(\mu(a^2+1)+x^2)^3}{x^2(\mu(a^2+1)+x^2)^3}\\
=&\frac{a((\mu(a^2+1)+x^2)^2+(\mu(a^2+1)+x^2)(4\mu a)^{1/3}x+(4\mu a)^{2/3}x^2)((4\mu a)^{1/3}x-\mu(a^2+1)-x^2)}{x^2(\mu(a^2+1)+x^2)^3}
\end{aligned}$$

For $\mu>\frac{a^2}{4(a^2+1)^3}$, then $(4\mu a)^{1/3}x-\mu(a^2+1)-x^2<0\Leftrightarrow\frac{dt(x,\mu)}{dx}<0$. For $\mu<\frac{a^2}{4(a^2+1)^3}$, and $x_{\rm lb}<x<x_{\rm ub}$, then $(4\mu a)^{1/3}x-\mu(a^2+1)-x^2>0\Leftrightarrow\frac{dt(x,\mu)}{dx}>0$, For $\mu<\frac{a^2}{4(a^2+1)^3}$, $x<x_{\rm lb}$ or $x>x_{\rm ub}$, $(4\mu a)^{1/3}x-\mu(a^2+1)-x^2<0\Leftrightarrow\frac{dt(x,\mu)}{dx}<0$.

We use the same argument as before to show that

$$x_{\rm lb}<\sqrt{\mu(a^2+1)}<x_{\rm ub}$$

(iv) Note that for $0<\mu<\frac{a^2}{4(a^2+1)^3}$

$$\frac{\partial t}{\partial\mu}=-a^2\frac{x^2-\mu(a^2+1)}{(\mu(a^2+1)+x^2)^3}\quad\text{and}\quad x_{\rm lb}<\sqrt{\mu(a^2+1)}<x_{\rm ub}$$

then $\frac{\partial t}{\partial\mu}\Big|_{x=x_{\rm lb}}>0$. Let $q(\mu)=t(x_{\rm lb},\mu)$, because $\frac{\partial t}{\partial x}\big|_{x=x_{\rm lb}}=0$, then

$$\frac{dq(\mu)}{d\mu}=\frac{dt(x_{\rm lb},\mu)}{d\mu}=\frac{\partial t}{\partial x}\Big|_{x=x_{\rm lb}}\frac{dx_{\rm lb}}{d\mu}+\frac{\partial t}{\partial\mu}\Big|_{x=x_{\rm lb}}=\frac{\partial t}{\partial\mu}\Big|_{x=x_{\rm lb}}>0$$

Note that $\mu=\frac{a^2}{4(a^2+1)^3}$, $x_{\rm ub}=x_{\rm lb}=\frac{(4\mu a)^{1/3}}{2}$, $t(\frac{(4\mu a)^{1/3}}{2},\frac{a^2}{4(a^2+1)^3})=\frac{a}{(4\mu a)^{1/3}}-1>0$. When $\mu<\left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2$, then $t(\sqrt{\mu(a^2+1)},\mu)<0$ by Theorem 6(ii). It implies that $q(\mu)<0$ when $\mu\to0^+$. By Theorem 6(iii), $q(\mu)=t(x_{\rm lb},\mu)<t(\sqrt{\mu(a^2+1)},\mu)<0$. Combining all of the theses, i.e.

$$\frac{dq(\mu)}{d\mu}>0,\quad\lim_{\mu\to0^+}q(\mu)<0,\quad q(\frac{a^2}{4(a^2+1)^3})>0$$

There exists a $\tau<\frac{a^2}{4(a^2+1)^3}$, $q(\tau)=0$. Such τ is the same as in Theorem 5(iv).

(v) We follow the same proof from the proof of Theorem 5(v).

(vi) By Theorem 6(v), $\forall\mu<\mu_0$, there exists three stationary points such that $0<x_\mu^{***}<x_{\rm lb}<x_\mu^{**}<x_{\rm ub}<x_\mu^*<a$. Because $\frac{dt(x;\mu)}{dx}\Big|_{x=x_{\rm lb}}=\frac{dt(x;\mu)}{dx}\Big|_{x=x_{\rm ub}}=0$, then

$$\frac{dt(x;\mu)}{dx}\Big|_{x=x_\mu^*}\neq0,\quad\frac{dt(x;\mu)}{dx}\Big|_{x=x_\mu^{**}}\neq0,\quad\frac{dt(x;\mu)}{dx}\Big|_{x=x_\mu^{***}}\neq0$$

By implicit function theorem [14], for solutions to equation $t(x;\mu) = 0$, there exists a unique continuously differentiable function such that $x = x(\mu)$ and satisfies $t(x(\mu), \mu) = 0$. Therefore,

$$\frac{dx}{d\mu} = -\frac{\partial t/\partial \mu}{\partial t/\partial x} = a^2 \frac{\frac{x^2-\mu(a^2+1)}{(\mu(a^2+1)+x^2)^3}}{-\frac{a}{x^2} + \frac{4\mu a^2 x}{(\mu(a^2+1)+x^2)^3}}$$

Therefore, by Theorem 6(iii)

$$\left.\frac{dx}{d\mu}\right|_{x=x_\mu^*} < 0 \quad \left.\frac{dx}{d\mu}\right|_{x=x_\mu^{***}} > 0$$

Because $0 < x_\mu^{***} < x_{\text{lb}} < x_\mu^{**} < x_{\text{ub}}$ and $\lim_{\mu\to 0^+} x_{\text{lb}} = \lim_{\mu\to 0^+} x_{\text{ub}} = 0$.

$$\lim_{\mu\to 0} x_\mu^{**} = \lim_{\mu\to 0} x_\mu^{***} = 0$$

Let us consider $t(a(1-c\mu), \mu)$ where $c = \frac{32}{a^2}$ and $\mu < \frac{1}{2c}$

$$\begin{aligned}
&t(a(1-c\mu);\mu) \\
=&\frac{a}{a(1-c\mu)} - \frac{\mu a^2}{(\mu(a^2+1)+a^2(1-c\mu)^2)^2} - 1 \\
=&\frac{c\mu}{1-c\mu} - \frac{\mu a^2}{(\mu(a^2+1)+a^2(1-c\mu)^2)^2} \\
\geq& c\mu - \frac{\mu a^2}{(a^2(1-c\mu)^2)^2} \\
\geq& c\mu - \frac{16}{a^2}\mu > 0
\end{aligned}$$

By Theorem 6(iii). It implies

$$a(1-c\mu) \leq x_\mu^*$$

taking $\mu \to 0^+$ on both side,

$$a = \lim_{\mu\to 0^+} a(1-c\mu) \leq \lim_{\mu\to 0^+} x_\mu^* \leq a$$

Hence, $\lim_{\mu\to 0} x_\mu^* = a$.

When $\mu = \tau$, because $t(x_{\text{lb}};\mu) = 0$ and $x_{\text{ub}} > \sqrt{\mu(a^2+1)} > x_{\text{lb}}$, $t(x;\mu)$ is increasing function between $[x_{\text{lb}}, x_{\text{ub}}]$ then $t(\sqrt{\mu(a^2+1)};\mu) > t(x_{\text{lb}};\mu) = 0$. Moreover, $t(\sqrt{\mu(a^2+1)},\mu)$, x_{lb} and x_μ^{**} are continuous function w.r.t μ, $\exists\delta > 0$ which is really small, such that $\mu = \tau - \delta$ and $t(\sqrt{\mu(a^2+1)},\mu) > 0$, $t(x_{\text{lb}},\mu) < 0$ (by Theorem 6(iv)) and $x_\mu^{**} > x_{\text{lb}}$, hence $\left.\frac{dx}{d\mu}\right|_{x=x_\mu^{**}} < 0$. It implies when μ decreases, then x_μ^{**} increases. This relation holds until $x_\mu^{**} = \sqrt{\mu(a^2+1)}$

$$\begin{aligned}
&t(x_\mu^{**},\mu) = t(\sqrt{\mu(a^2+1)},\mu) = 0 \\
\Rightarrow&\mu = \left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2
\end{aligned}$$

and $\sqrt{\mu(a^2+1)} = \frac{a(\sqrt{a^2+1}-a)}{2\sqrt{a^2+1}}$. Note that when $\mu < \left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2$, $t(\sqrt{\mu(a^2+1)},\mu) < 0$, it implies that $x_\mu^{**} > \sqrt{\mu(a^2+1)}$ and $\left.\frac{dx}{d\mu}\right|_{x=x_\mu^{**}} > 0$, thus decreasing μ leads to decreasing x_μ^{**}. We can conclude

$$\max_{\mu\leq\tau} x_\mu^{**} \leq \frac{a(\sqrt{a^2+1}-a)}{2\sqrt{a^2+1}}$$

Note that $\forall \mu$ s.t. $\left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2 < \mu < \tau$, $x_\mu^{**} < \left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2$, so $t(\left(\frac{a(\sqrt{a^2+1}-a)}{2(a^2+1)}\right)^2, \mu) \geq 0$.

Note that when $\mu > \frac{a^2}{a^2+1}$, i.e. $(x_\mu^*)^2 \geq \mu(a^2+1)$ then

$$\left.\frac{dx}{d\mu}\right|_{x=x_\mu^*} > 0$$

It implies that when μ decreases, x_μ^* also decreases. It holds true until $x_\mu^* = \sqrt{\mu(a^2+1)}$. The same analysis can be applied to x_μ^* like above, we can conclude that

$$\min_\tau x_\mu^* = \frac{a(\sqrt{a^2+1}+a)}{2\sqrt{a^2+1}}$$

Hence

$$\max_{\mu \leq \tau} x_\mu^{**} \leq \frac{a(\sqrt{a^2+1}-a)}{2\sqrt{a^2+1}} < \frac{a(\sqrt{a^2+1}+a)}{2\sqrt{a^2+1}} \leq \min_{\mu>0} x_\mu^*$$

□

E.4 Proof of Theorem 7, 8 and 9

Proof. The proof is similar to the proof of Theorem 5 and Theorem 6. □

E.5 Proof of Lemma 1

Proof.

$$\nabla^2 g_\mu(x,y) = \begin{pmatrix} \mu + y^2 & 2xy \\ 2xy & \mu(a^2+1)+x^2 \end{pmatrix}$$

Let $\lambda_1(\nabla^2 g_\mu(x,y)), \lambda_2(\nabla^2 g_\mu(x,y))$ be the eigenvalue of matrix $\nabla^2 g_\mu(x,y)$, then

$$\begin{aligned} &\lambda_1(\nabla^2 g_\mu(x,y)) + \lambda_2(\nabla^2 g_\mu(x,y)) \\ =&\operatorname{Tr}(\nabla^2 g_\mu(x,y)) = \mu + y^2 + \mu(a^2+1) + x^2 > 0 \end{aligned}$$

Now we calculate the product of eigenvalue

$$\begin{aligned} &\lambda_1(\nabla^2 g_\mu(x,y)) \cdot \lambda_2(\nabla^2 g_\mu(W)) \\ =&\det(\nabla^2 g_\mu(W)) \\ =&(\mu+y^2)(\mu(a^2+1)+x^2) - 4x^2y^2 \\ =&\frac{\mu a}{x}\frac{\mu a}{y} - 4x^2y^2 > 0 \\ \Leftrightarrow&(\frac{a\mu}{2})^{2/3} > xy \\ \Leftrightarrow&(\frac{a\mu}{2})^{2/3} > \frac{a\mu}{y^2+\mu} y \\ \Leftrightarrow&y + \frac{\mu}{y} > (4a\mu)^{1/3} \end{aligned}$$

Note that for $(x_\mu^*, y_\mu^*), (x_\mu^{***}, y_\mu^{***})$, they satisfy (11a) and (11b), this fact is used in third equality and second "⇔". By (32b), we know $\lambda_1(\nabla^2 g_\mu(x,y)) \cdot \lambda_2(\nabla^2 g_\mu(x,y)) > 0$ for $(x_\mu^*, y_\mu^*), (x_\mu^{***}, y_\mu^{***})$, and $\lambda_1(\nabla^2 g_\mu(x,y)) \cdot \lambda_2(\nabla^2 g_\mu(x,y)) < 0$ for (x_μ^{**}, y_μ^{**}), then

$$\begin{aligned} &\lambda_1(\nabla^2 g_\mu(x,y)) > 0, \lambda_2(\nabla^2 g_\mu(x,y)) > 0 \qquad \text{for } (x_\mu^*, y_\mu^*), (x_\mu^{***}, y_\mu^{***}) \\ &\lambda_1(\nabla^2 g_\mu(x,y)) < 0 \text{ or } \lambda_2(\nabla^2 g_\mu(x,y)) < 0 \qquad \text{for } (x_\mu^{**}, y_\mu^{**}) \end{aligned}$$

and

$$\nabla g_\mu(x,y) = 0$$

Then $(x_\mu^*, y_\mu^*), (x_\mu^{***}, y_\mu^{***})$ are locally minima, (x_μ^{**}, y_μ^{**}) is saddle point for $g_\mu(W)$. □

E.6 Proof of Lemma 2

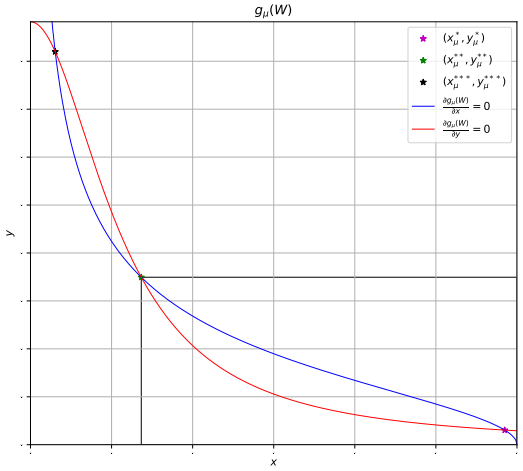


Figure 8: Stationary points when $\mu < \tau$

Proof. Let us define the functions as below

$$\begin{cases} y_{\mu 1}(x) = \sqrt{\mu(\dfrac{a-x}{x})} & 0 < x \le a \quad (37a) \\ y_{\mu 2}(x) = \dfrac{\mu a}{\mu(a^2+1)+x^2} & 0 < x \le a \quad (37b) \end{cases}$$

$$\begin{cases} x_{\mu 1}(y) = \dfrac{\mu a}{y^2+\mu} & 0 < y < \frac{a}{a^2+1} \quad (38a) \\ x_{\mu 2}(y) = \sqrt{\mu(\dfrac{a}{y} - (a^2+1))} & 0 < y < \frac{a}{a^2+1} \quad (38b) \end{cases}$$

with simple calculations,

$$y_{\mu 1} \ge y_{\mu 2} \Leftrightarrow t(x;\mu) \ge 0 \Leftrightarrow x \in (0, x_\mu^{***}] \cup [x_\mu^{**}, x_\mu^*]$$

and

$$x_{\mu 1} \ge x_{\mu 2} \Leftrightarrow r(y;\mu) \le 0 \Leftrightarrow y \in [y_\mu^*, y_\mu^{**}] \cup [y_\mu^{***}, \frac{a}{a^2+1})$$

Here we divide B_μ into three parts, $C_{\mu 1}, C_{\mu 2}, C_{\mu 3}$

$$C_{\mu 1} = \{(x,y) | x_\mu^{**} < x \le x_\mu^*, y_{\mu 1} < y < y_\mu^{**}\} \cup \{(x,y) | x_\mu^* < x \le a, y_{\mu 2} < y < y_\mu^{**}\} \quad (39)$$
$$C_{\mu 2} = \{(x,y) | x_\mu^{**} < x \le x_\mu^*, 0 \le y < y_{\mu 2}\} \cup \{(x,y) | x_\mu^* < x \le a, 0 \le y < y_{\mu 1}\} \quad (40)$$
$$C_{\mu 3} = \{(x,y) | x_\mu^{**} < x \le x_\mu^*, y_{\mu 2} \le y \le y_{\mu 1}\} \cup \{(x,y) | x_\mu^* < x \le a, y_{\mu 1} \le y \le y_{\mu 2}\} \quad (41)$$

Also note that

$$\forall (x,y) \in C_{\mu 1} \Rightarrow \frac{\partial g_\mu(x,y)}{\partial x} > 0, \frac{\partial g_\mu(x,y)}{\partial y} > 0$$
$$\forall (x,y) \in C_{\mu 2} \Rightarrow \frac{\partial g_\mu(x,y)}{\partial x} < 0, \frac{\partial g_\mu(x,y)}{\partial y} < 0$$

The gradient flow follows

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = -\begin{pmatrix} \frac{\partial g_\mu(x(t),y(t))}{\partial x} \\ \frac{\partial g_\mu(x(t),y(t))}{\partial y} \end{pmatrix} = -\nabla g_\mu(x(t), y(t))$$

then

$$\forall (x,y) \in C_{\mu 1} \Rightarrow \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} < 0, \quad \|\nabla g_\mu\| > 0 \tag{42}$$

$$\forall (x,y) \in C_{\mu 2} \Rightarrow \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} > 0, \quad \|\nabla g_\mu\| > 0 \tag{43}$$

Note that $\|\nabla g_\mu\|$ is not diminishing and bounded away from 0. Let us consider the $(x(0), y(0)) \in C_{\mu 1}$, since $\nabla g_\mu(x,y) \neq 0$, $-\nabla g_\mu(x,y) < 0$ in (42) and boundness of $C_{\mu 1}$, it implies there exists a finite $t_0 > 0$ such that

$$(x(t_0), y(t_0)) \in \partial C_{\mu 1}, (x(t), y(t)) \in C_{\mu 1} \text{ for } 0 \leq t < t_0$$

where $\partial C_{\mu 1}$ is defined as

$$\partial C_{\mu 1} = \{(x,y) | x_\mu^{**} < x \leq x_\mu^*, y = y_{\mu 1}\} \cup \{(x,y) | x_\mu^* < x \leq a, y = y_{\mu 2}\} \subseteq C_{\mu 3}$$

For the same reason, if $(x(0), y(0)) \in C_{\mu 2}$, there exists a finite time $t_1 > 0$,

$$(x(t_0), y(t_0)) \in \partial C_{\mu 2}, (x(t), y(t)) \in C_{\mu 2} \text{ for } 0 \leq t < t_1$$

where $\partial C_{\mu 2}$ is defined as

$$\partial C_{\mu 2} = \{(x,y) | x_\mu^{**} < x \leq x_\mu^*, y = y_{\mu 2}\} \cup \{(x,y) | x_\mu^* < x \leq a, y = y_{\mu 1}\} \subseteq C_{\mu 3}$$

then by lemma 7, $\lim_{t\to\infty}(x(t), y(t)) = (x_\mu^*, y_\mu^*)$. □

E.7 Proof of Lemma 3

Proof. This is just a result of the Theorem 5. □

E.8 Proof of Lemma 5

Proof. Note that

$$\nabla^2 g_\mu(W) = \begin{pmatrix} \mu + y^2 & 2xy \\ 2xy & \mu(a^2+1) + x^2 \end{pmatrix} = \begin{pmatrix} \mu & 0 \\ 0 & \mu(a^2+1) \end{pmatrix} + \begin{pmatrix} y^2 & 2xy \\ 2xy & x^2 \end{pmatrix}$$

Let $\|\cdot\|_{\text{op}}$ is the spectral norm, and it satisfies triangle inequality

$$\begin{aligned} \left\|\nabla^2 g_\mu(W)\right\|_{\text{op}} \leq & \left\| \begin{pmatrix} \mu & 0 \\ 0 & \mu(a^2+1) \end{pmatrix} \right\|_{\text{op}} + \left\| \begin{pmatrix} y^2 & 2xy \\ 2xy & x^2 \end{pmatrix} \right\|_{\text{op}} \\ = & \mu(a^2+1) + \left\| \begin{pmatrix} y^2 & 2xy \\ 2xy & x^2 \end{pmatrix} \right\|_{\text{op}} \end{aligned}$$

The spectral norm of the second term in area A is bounded by

$$\max_{(x,y)\in A} \frac{(x^2+y^2) + \sqrt{(x^2+y^2)^2 + 12x^2y^2}}{2} \leq \frac{2a^2 + \sqrt{4a^4 + 12a^4}}{2} = 3a^2$$

We use $x^2 \leq a^2, y^2 \leq a^2$ in the inequality. Therefore,

$$\left\|\nabla^2 g_\mu(W)\right\|_{\text{op}} \leq 3a^2 + \mu(a^2+1)$$

Also, according to [5, 33], for any f, if $\nabla^2 f$ exists, then f is L smooth if and only if $|\nabla^2 f|_{\text{op}} \leq L$. With this, we conclude the proof. □

E.9 Proof of Lemma 7

Proof. First we prove $\forall t \geq 0, (x(t), y(t)) \in C_{\mu 3}$, because if $(x(t), y(t)) \notin C_{\mu 3}$, then there exists a finite t such that

$$(x(t), y(t)) \in \partial C_{\mu 3}$$

where $\partial C_{\mu 3}$ is the boundary of $C_{\mu 3}$, defined as

$$\partial C_{\mu 3} = \{(x,y) | y = y_{\mu 1}(x) \text{ or } y = y_{\mu 2}(x), x_\mu^{**} < x \leq a\}$$

W.L.O.G, let us assume $(x(0), y(0)) \in \partial C_{\mu 3}$ and $(x(0), y(0)) \neq (x_\mu^*, y_\mu^*)$. Here are four different cases,

$$\nabla g_\mu(x(t), y(t)) = \begin{cases} \begin{pmatrix} = 0 \\ > 0 \end{pmatrix} & \text{if } y(0) = y_{\mu 1}(x(0)), x_\mu^{**} < x(0) < x_\mu^* \\ \begin{pmatrix} = 0 \\ < 0 \end{pmatrix} & \text{if } y(0) = y_{\mu 1}(x(0)), x_\mu^* < x(0) \leq a \\ \begin{pmatrix} < 0 \\ = 0 \end{pmatrix} & \text{if } y(0) = y_{\mu 2}(x(0)), x_\mu^{**} < x(0) < x_\mu^* \\ \begin{pmatrix} > 0 \\ = 0 \end{pmatrix} & \text{if } y(0) = y_{\mu 2}(x(0)), x_\mu^* < x(0) \leq a \end{cases}$$

This indicates that $-\nabla g_\mu(x(t), y(t))$ are pointing to the interior of $C_{\mu 3}$, then $(x(t), y(t))$ can not escape $C_{\mu 3}$. Here we can focus our attention in $C_{\mu 3}$, because $\forall t \geq 0, (x(t), y(t)) \in C_{\mu 3}$. For Algorithm 1,

$$\frac{df(\boldsymbol{z}_t)}{dt} = \nabla f(\boldsymbol{z}_t)\dot{\boldsymbol{z}}_t = -\|\nabla f(\boldsymbol{z}_t)\|_2^2$$

In our setting, $\forall (x, y) \in C_{\mu 3}$

$$\begin{cases} \nabla g_\mu(x, y) \neq 0 & (x, y) \neq (x_\mu^*, y_\mu^*) \\ \nabla g_\mu(x, y) = 0 & (x, y) = (x_\mu^*, y_\mu^*) \end{cases}$$

so

$$\frac{dg_\mu(x(t), y(t))}{dt} = \begin{cases} -\|\nabla g_\mu\|_2^2 < 0 & (x, y) \neq (x_\mu^*, y_\mu^*) \\ -\|\nabla g_\mu\|_2^2 = 0 & (x, y) = (x_\mu^*, y_\mu^*) \end{cases}$$

Plus, (x_μ^*, y_μ^*) is the unique stationary point of $g_\mu(W)$ in $C_{\mu 3}$. By lemma 8

$$g_\mu(x, y) > g_\mu(x_\mu^*, y_\mu^*) \quad (x, y) \neq (x_\mu^*, y_\mu^*)$$

By Lyapunov asymptotic stability theorem [28], and applying it to gradient flow for $g_\mu(x, y)$ in $C_{\mu 3}$, we can conclude $\lim_{t\to\infty}(x(t), y(t)) = (x_\mu^*, y_\mu^*)$. □

E.10 Proof of Lemma 8

Proof. For any $(x, y) \in C_{\mu 3}$ in 41, and $(x, y) \neq (x_\mu^*, y_\mu^*)$, in Algorithm 7, W.L.O.G, we can assume $x \in (x_\mu^{**}, x_\mu^*)$, the analysis details can also be applied to $x \in (x_\mu^*, a)$. It is obvious that $\tilde{x}_j < \tilde{x}_{j+1}$ and $\tilde{y}_{j+1} < \tilde{y}_j$. Also, $\lim_{j\to\infty}(\tilde{x}_j, \tilde{y}_j) = (x_\mu^*, y_\mu^*)$. Otherwise either $\tilde{x}_j \neq x_\mu^*$ or $\tilde{y}_j \neq y_\mu^*$ hold, Algorithm 7 continues until $\lim_{j\to\infty}(\tilde{x}_j, \tilde{y}_j) = \lim_{j\to\infty}(y_{\mu 2}(\tilde{y}_j), x_{\mu 1}(\tilde{x}_j))$, i.e. $(\tilde{x}_j, \tilde{y}_j)$ converges to (x_μ^*, y_μ^*).

Moreover, note that for any $j = 0, 1, \ldots$

$$g_\mu(\tilde{x}_{j-1}, \tilde{y}_{j-1}) > g_\mu(\tilde{x}_{j-1}, \tilde{y}_j) > g_\mu(\tilde{x}_j, \tilde{y}_j)$$

Because

$$g_\mu(\tilde{x}_{j-1}, \tilde{y}_{j-1}) - g_\mu(\tilde{x}_{j-1}, \tilde{y}_j) = \frac{\partial g_\mu(\tilde{x}_{j-1}, \tilde{y})}{\partial y}(\tilde{y}_{j-1} - \tilde{y}_j) \quad \text{where } \tilde{y} \in (\tilde{y}_j, \tilde{y}_{j-1})$$

Note that

$$\frac{\partial g_\mu(\tilde{x}_{j-1}, \tilde{y})}{\partial y} > 0 \Rightarrow g_\mu(\tilde{x}_{j-1}, \tilde{y}_{j-1}) > g_\mu(\tilde{x}_{j-1}, \tilde{y}_j)$$

By the same reason,

$$g_\mu(\tilde{x}_{j-1}, \tilde{y}_j) > g_\mu(\tilde{x}_j, \tilde{y}_j)$$

By Lemma 1, (x_μ^*, y_μ^*) is local minima, and there exists a $r_\mu > 0$ and any $\{(x, y) \mid \|(x, y) - (x_\mu^*, y_\mu^*)\|_2 \leq r_\mu\}, g_\mu(x, y) > g_\mu(x_\mu^*, y_\mu^*)$ Since $\lim_{j\to\infty}(\tilde{x}_j, \tilde{y}_j) = (x_\mu^*, y_\mu^*)$, there exists a $J > 0$ such that $\forall j > J$, $\|(\tilde{x}_j, \tilde{y}_j) - (x_\mu^*, y_\mu^*)\|_2 \leq r_\mu$, combining them all

$$g_\mu(x, y) > g_\mu(\tilde{x}_j, \tilde{y}_j) > g_\mu(x_\mu^*, y_\mu^*)$$

□

Algorithm 7: Path goes to (x^*_μ, y^*_μ)

Input: $(x, y) \in C_{\mu 3}, x_{\mu 1}(y), y_{\mu 2}(x)$ as (38a),(37b)
Output: $\{(\tilde{x}_j, \tilde{y}_j)\}_{j=0}^{\infty}$

$(\tilde{x}_0, \tilde{y}_0) \leftarrow (x, y)$
for $j = 1, 2, \ldots$ **do**
  $\tilde{y}_j \leftarrow y_{\mu 2}(\tilde{x}_{j-1})$
  $\tilde{x}_j \leftarrow x_{\mu 1}(\tilde{y}_{j-1})$
end

E.11 Proof of Lemma 4

Proof. From the proof of Theorem 1, any any scheduling for μ_k satisfies following will do the job

$$(2/a)^{2/3}\mu_{k-1}^{4/3} \leq \mu_k < \mu_{k-1}$$

Note that in Algorithm 4, we have $\widehat{a} = \sqrt{4(\mu_0 + \varepsilon)} < a$, then it is obvious

$$(2/a)^{2/3}\mu_{k-1}^{4/3} < (2/\widehat{a})^{2/3}\mu_{k-1}^{4/3}$$

The same analysis for Theorem 1 can be applied here. □

E.12 Proof of Lemma 6

Proof. By the Theorem 3 and Lemma 5 and the fact that $A^1_{\mu,\epsilon}$ is μ-stationary point region, we use the same argument as proof of Lemma 7 to demonstrate the gradient descent will never go to $A^2_{\mu,\epsilon}$. □

E.13 Proof of Lemma 9

Proof. By Theorem 9(iv)

$$\max_{\mu \leq \tau_\beta} x^{**}_{\mu,\beta} \leq \min_{\mu > 0} x^*_{\mu,\beta}$$

We also know from the proof of Corollary 3, $x^{**}_{\mu,\epsilon} < x^{**}_{\mu,\beta}$ and $x^*_{\mu,\beta} < x^*_{\mu,\epsilon}$. Consequently,

$$\max_{\mu \leq \tau_\beta} x^{**}_{\mu,\epsilon} \leq \min_{\mu > 0} x^*_{\mu,\epsilon}$$

Because $\tau_\beta > \tau$, so

$$\max_{\mu \leq \tau} x^{**}_{\mu,\epsilon} \leq \max_{\mu \leq \tau_\beta} x^{**}_{\mu,\epsilon} \leq \min_{\mu > 0} x^*_{\mu,\epsilon}$$

□

E.14 Proof of Corollary 1

Proof. Note that

$$\frac{a^2}{4(a^2+1)^3} \leq \frac{1}{27} \quad a > 0$$

when $a > \sqrt{\frac{5}{27}}$, then $\frac{a^2}{4} > \mu_0 = \frac{1}{27} \geq \frac{a^2}{4(a^2+1)^3}$, it satisfies condition in Lemma 4, we obtain the same result. □

E.15 Proof of Corollary 2

Proof. Use Theorem 5(vi) and Theorem 6(vi). □

E.16 Proof of Corollary 3

Proof. It is easy to know that

$$r_\beta(y;\mu) > r_\epsilon(y;\mu) > r(y;\mu)$$

and

$$t_\beta(x;\mu) < t_\epsilon(x;\mu) < t(x;\mu)$$

and when $\mu < \tau$, there are three solutions to $r(y;\mu) = 0$ by Theorem 5. Also, we know from Theorem 7, 8

$$\lim_{y\to 0^+} r_\epsilon(y;\mu) = \infty \qquad \lim_{y\to 0^+} r_\beta(y;\mu) = \infty$$

Note that when $\left(\frac{1+\beta}{1-\beta}\right)^2 \le a^2 + 1$

$$r_\beta(\sqrt{\mu};\mu) = \frac{a(1+\beta)}{\sqrt{\mu}} - (a^2+1) - \frac{a^2(1-\beta)^2}{4\mu} \le 0 \quad \forall \mu > 0$$

Therefore,

$$0 \ge r_\beta(\sqrt{\mu};\mu) > r_\epsilon(\sqrt{\mu};\mu) > r(\sqrt{\mu};\mu)$$

Also, we know that for y_{ub} defined in Theorem 5(iii), we know $r(y_{\text{ub}};\mu) > 0$ from Theorem 5(iv). Therefore,

$$r_\beta(y_{\text{ub}};\mu) > r_\epsilon(y_{\text{ub}};\mu) > r(y_{\text{ub}};\mu) > 0$$

Besides, $\sqrt{\mu} < y_{\text{ub}}$. By monotonicity of $r_\beta(y;\mu)$ and $r_\epsilon(y;\mu)$ from the Theorem 7(ii) and Theorem 8(ii), it implies that there are at least two solutions to $r_\beta(y;\mu)$ and $r_\epsilon(y;\mu)$. From the geometry of $r_\beta(y;\mu), r_\epsilon(y;\mu), r(y;\mu)$ and $t_\beta(x;\mu), t_\epsilon(x;\mu), t(x;\mu)$, it is trivial to know that $x^*_{\mu,\epsilon} \le x^*_\mu$, $y^*_{\mu,\epsilon} \ge y^*_\mu$, $x^{**}_{\mu,\epsilon} \ge x^{**}_\mu$, $y^*_{\mu,\epsilon} \le y^{**}_\mu$.

Finally, for every point $(x,y) \in A^1_{\mu,\epsilon}$, there exists a pair ϵ_1, ϵ_2, each satisfying $|\epsilon_1| \le \epsilon$ and $|\epsilon_2| \le \epsilon$, such that (x,y) is the solution to

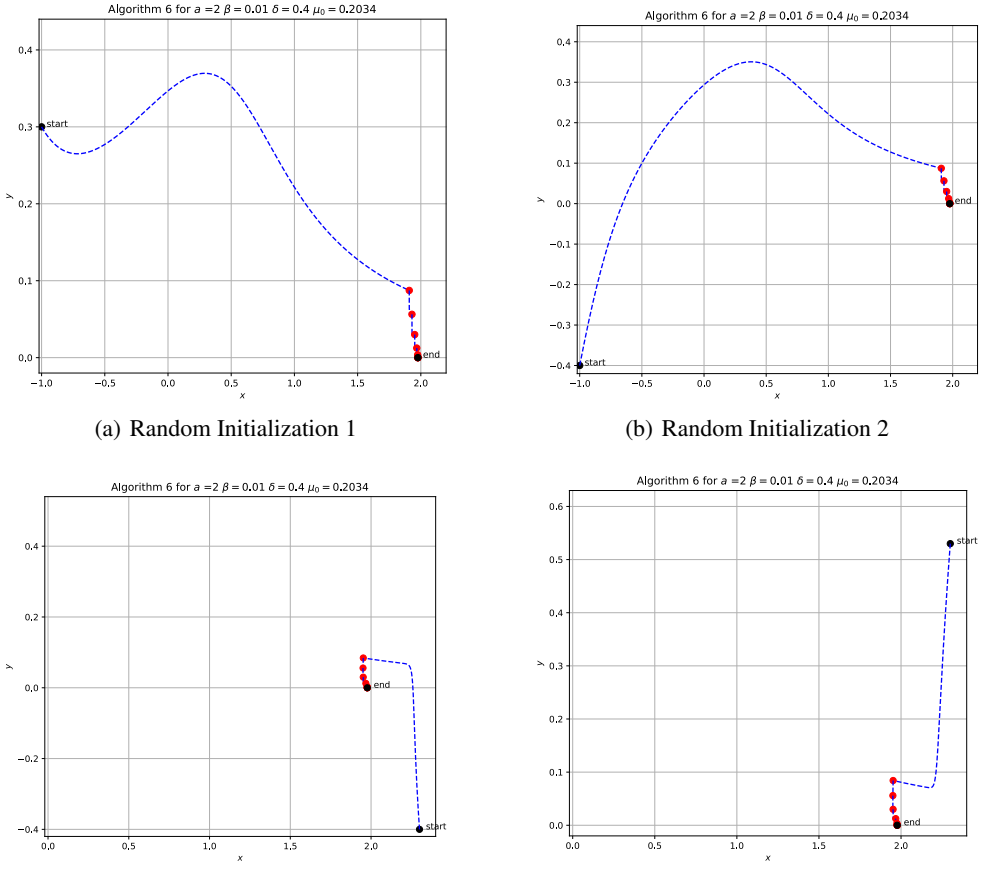
$$x = \frac{\mu a + \epsilon_1}{\mu + y^2} \qquad y = \frac{\mu a + \epsilon_2}{x^2 + \mu(a^2+1)}$$

We can repeat the same analysis above to show that $x^*_{\mu,\epsilon} \le x$, $y^*_{\mu,\epsilon} \ge y$. Applying the same logic to $\forall (x,y) \in A^2_{\mu,\epsilon}$, we find $x^{**}_{\mu,\epsilon} \ge x$, $y^*_{\mu,\epsilon} \le y$. Thus, (x^*_μ, y^*_μ) is the extreme point of $A^1_{\mu,\epsilon}$ and (x^{**}_μ, y^{**}_μ) is the extreme point of $A^2_{\mu,\epsilon}$, we get the results. □

F Experiments Details

In this section, we present experiments to validate the global convergence of Algorithm 6. Our goal is twofold: First, we aim to demonstrate that irrespective of the starting point, Algorithm 6 using gradient descent consistently returns the global minimum. Second, we contrast our updating scheme for μ_k, ϵ_k as prescribed in Algorithm 6 with an arbitrary updating scheme for μ_k, ϵ_k. This comparison illustrates how inappropriate setting of parameters in gradient descent could lead to incorrect solutions.

F.1 Random Initialization Converges to Global Optimum



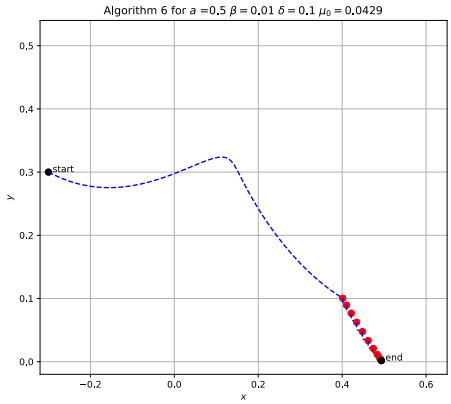
(a) Random Initialization 1

(b) Random Initialization 2

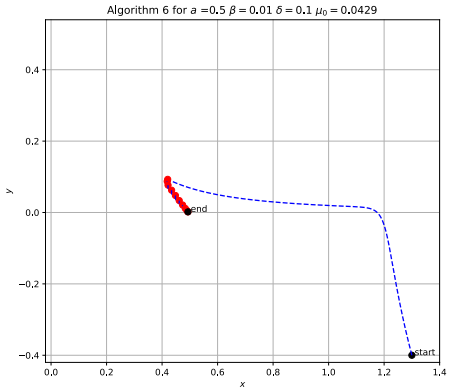
(c) Random Initialization 3

(d) Random Initialization 4

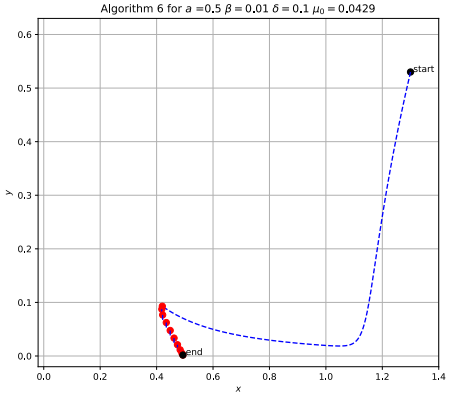
Figure 9: Trajectory of the gradient descent path with the different initializations for $a = 2$. We observe that regardless of the initialization, Algorithm 6 always converges to the global minimum. Initial $\mu_0 = \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$



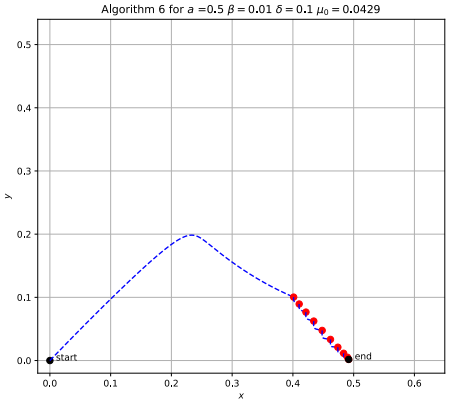
(a) Random Initialization 1



(b) Random Initialization 2



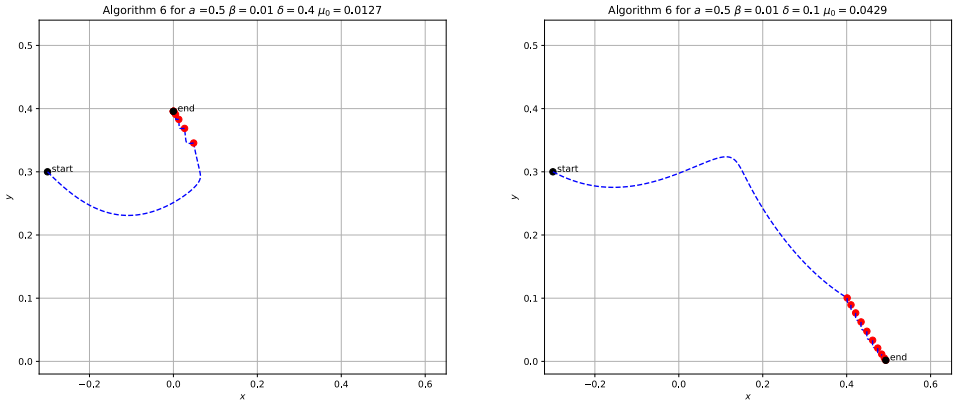
(c) Random Initialization 3



(d) Random Initialization 4

Figure 10: Trajectory of the gradient descent path with the different initializations for $a = 0.5$. We observe that regardless of the initialization, Algorithm 6 always converges to the global minimum. Initial $\mu_0 = \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$

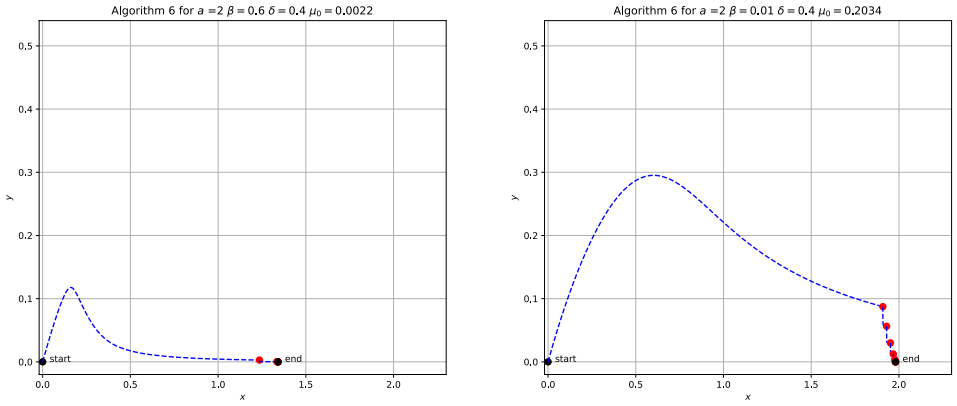
F.2 Wrong Specification of δ Leads to Spurious Local Optimial



(a) $\delta = 0.4$ (b) $\delta = 0.1$

Figure 11: Trajectory of the gradient descent path for two difference δ. Left: β violates requirement $\left(\frac{1+\beta}{1-\beta}\right)^2 \leq (1-\delta)(a^2+1)$ in Theorem 4, leading to spurious local minimum. Right: β follows requirement $\left(\frac{1+\beta}{1-\beta}\right)^2 \leq (1-\delta)(a^2+1)$ in Theorem 4, leading to global minimum. Initial $\mu_0 = \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$

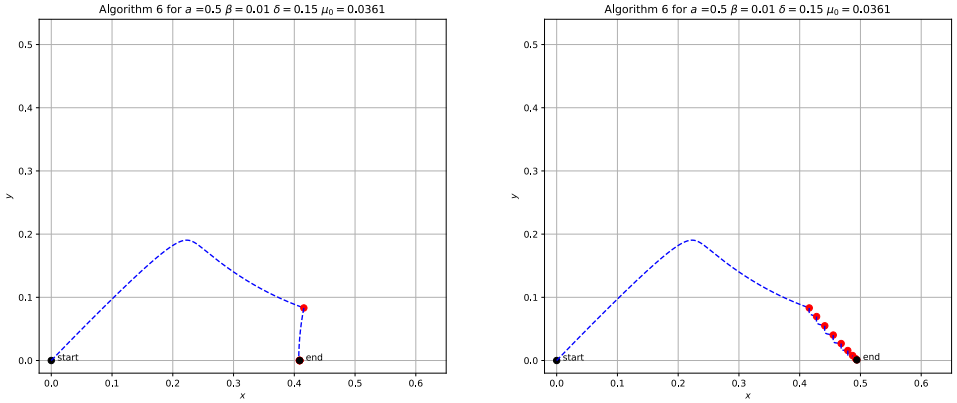
F.3 Wrong Specification of β Leads to Incorrect Solution



(a) $\beta = 0.6$ (b) $\delta = 0.01$

Figure 12: Trajectory of the gradient descent path for two difference β. Left: β violates requirement $\left(\frac{1+\beta}{1-\beta}\right)^2 \leq (1-\delta)(a^2+1)$ in Theorem 4, leading to incorrect solution. Right: β follows requirement $\left(\frac{1+\beta}{1-\beta}\right)^2 \leq (1-\delta)(a^2+1)$ in Theorem 4, leading to global minimum. Initial $\mu_0 = \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$

F.4 Faster decrease of μ_k Leads to Incorrect Solution



(a) Bad scheduling (b) Good scheduling

Figure 13: Trajectory of the gradient descent path for two difference update rules for μ_k with the same initialization. Left: "Bad scheduling" uses a faster-decreasing scheme for μ_k, leading to an incorrect solution, even a non-local optimal solution. Right: "Good scheduling" follows updating rule for μ_k in Algorithm 6, leading to the global minimum. Initial $\mu_0 = \frac{a^2}{4}\frac{(1-\delta)^3(1-\beta)^4}{(1+\beta)^2}$