521 # A   Appendix

522 ## A.1   Proof of Theorem 4.1

523 *Proof.* To begin with, let us consider Lv.1 GP:

$$\boldsymbol{\theta}_t^{(1)} = \boldsymbol{\theta}_t - \eta\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(0)}) = \boldsymbol{\theta}_t - \eta\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$$
$$\boldsymbol{\phi}_t^{(1)} = \boldsymbol{\phi}_t - \eta\nabla_{\boldsymbol{\phi}}g(\boldsymbol{\theta}_t^{(0)}, \boldsymbol{\phi}_t) = \boldsymbol{\phi}_t - \eta\nabla_{\boldsymbol{\phi}}g(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$$

524 The differences between $(\boldsymbol{\theta}_t^{(1)}, \boldsymbol{\phi}_t^{(1)})$ and $(\boldsymbol{\theta}_t^{(0)}, \boldsymbol{\phi}_t^{(0)})$ are:

$$\|\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(0)}\| = \eta\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)\|,$$
$$\|\boldsymbol{\phi}_t^{(1)} - \boldsymbol{\phi}_t^{(0)}\| = \eta\|\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)\|.$$

525 Recall our definition of $\Delta_{\max}$ we have:

$$\|\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(0)}\| + \|\boldsymbol{\phi}_t^{(1)} - \boldsymbol{\phi}_t^{(0)}\| \le \eta\Delta_{\max} \tag{9}$$

526 Then, with $\boldsymbol{\omega}_t = [\boldsymbol{\theta}_t, \boldsymbol{\phi}_t]^T$ and $\boldsymbol{\omega}_t^{(k)} = [\boldsymbol{\theta}_t^{(k)}, \boldsymbol{\phi}_t^{(k)}]^T$, the differences between Lv.2 agents and Lv.1
527 agents are:

$$\|\boldsymbol{\theta}_t^{(2)} - \boldsymbol{\theta}_t^{(1)}\| = \eta\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(1)}) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(0)})\|$$
$$\le \eta L_{\theta\phi}\|\boldsymbol{\phi}_t^{(1)} - \boldsymbol{\phi}_t^{(0)}\|,$$
$$\|\boldsymbol{\phi}_t^{(2)} - \boldsymbol{\phi}_t^{(1)}\| = \eta\|\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(1)}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(0)}, \boldsymbol{\phi}_t)\|$$
$$\le \eta L_{\phi\theta}\|\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^{(0)}\|.$$

528 Recall that $L := \max\{L_{\theta\theta}, L_{\theta\phi}, L_{\phi\theta}, L_{\phi\phi}\}$, using Equation (9) we have:

$$\|\boldsymbol{\theta}_t^{(2)} - \boldsymbol{\theta}_t^{(1)}\| + \|\boldsymbol{\phi}_t^{(2)} - \boldsymbol{\phi}_t^{(1)}\| \le \eta^2 L\Delta_{\max} \tag{10}$$

529 Similarly, we can derive the differences between Lv.3 and Lv.2 agents:

$$\|\boldsymbol{\theta}_t^{(3)} - \boldsymbol{\theta}_t^{(2)}\| = \eta\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(2)}) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(1)})\|$$
$$\le \eta L_{\theta\phi}\|\boldsymbol{\phi}_t^{(2)} - \boldsymbol{\phi}_t^{(1)}\|$$
$$\|\boldsymbol{\phi}_t^{(3)} - \boldsymbol{\phi}_t^{(2)}\| = \eta\|\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(1)}, \boldsymbol{\phi}_t)\|$$
$$\le \eta L_{\phi\theta}\|\boldsymbol{\theta}_t^{(2)} - \boldsymbol{\theta}_t^{(1)}\|$$
$$\|\boldsymbol{\theta}_t^{(3)} - \boldsymbol{\theta}_t^{(2)}\| + \|\boldsymbol{\phi}_t^{(3)} - \boldsymbol{\phi}_t^{(2)}\| \le \eta^3 L^2\Delta_{\max} \tag{11}$$

530 Consequently, the difference between any two consecutive states k and $k-1$ are upper bounded by:

$$\|\boldsymbol{\theta}_t^{(k)} - \boldsymbol{\theta}_t^{(k-1)}\| = \eta\|\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(k-1)}) - \nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(k-2)})\|$$
$$\le \eta L_{\theta\phi}\|\boldsymbol{\phi}_t^{(k-1)} - \boldsymbol{\phi}_t^{(k-2)}\|$$
$$\|\boldsymbol{\phi}_t^{(k)} - \boldsymbol{\phi}_t^{(k-1)}\| = \eta\|\nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(k-1)}, \boldsymbol{\phi}_t) - \nabla_{\boldsymbol{\phi}}f(\boldsymbol{\theta}_t^{(k-2)}, \boldsymbol{\phi}_t)\|$$
$$\le \eta L_{\phi\theta}\|\boldsymbol{\theta}_t^{(k-1)} - \boldsymbol{\theta}_t^{(k-2)}\|$$
$$\|\boldsymbol{\theta}_t^{(k)} - \boldsymbol{\theta}_t^{(k-1)}\| + \|\boldsymbol{\phi}_t^{(k)} - \boldsymbol{\phi}_t^{(k-1)}\| \le \eta \cdot (\eta L)^{(k-1)}\Delta_{\max} \tag{12}$$

531 Since $\|\boldsymbol{\omega}_t^{(k)} - \boldsymbol{\omega}_t^{(k-1)}\| \le \|\boldsymbol{\theta}_t^{(k)} - \boldsymbol{\theta}_t^{(k-1)}\| + \|\boldsymbol{\phi}_t^{(k)} - \boldsymbol{\phi}_t^{(k-1)}\|$ we have:

$$\|\boldsymbol{\omega}_t^{(k)} - \boldsymbol{\omega}_t^{(k-1)}\| \le \eta \cdot (\eta L)^{(k-1)}\Delta_{\max} \tag{13}$$

15

532 Suppose $\eta < (2L)^{-1}$, such that the difference between any two consecutive states is a contraction,
533 then we consider the difference, $\|\boldsymbol{\omega}_t^{(a)} - \boldsymbol{\omega}_t^{(b)}\|$, where $a > b > 0$. We can rewrite it as:

$$
\begin{aligned}
\|\boldsymbol{\omega}_t^{(a)} - \boldsymbol{\omega}_t^{(b)}\| &= \left\| \sum_{i=b+1}^{a} \boldsymbol{\omega}_t^{(i)} - \boldsymbol{\omega}_t^{(i-1)} \right\| \\
&\leq \sum_{i=b+1}^{a} \left\| \boldsymbol{\omega}_t^{(i)} - \boldsymbol{\omega}_t^{(i-1)} \right\| \\
&\leq \sum_{i=b+1}^{a} \eta \cdot (\eta L)^{(i-1)} \Delta_{\max} \\
&\leq \eta \Delta_{\max} \cdot \left[(\eta L)^{(b)} + \cdots + (\eta L)^{(a-1)} \right] \\
&\leq \eta \Delta_{\max} \cdot (\eta L)^{(b-1)} \\
&\leq \eta^b L^{(b-1)} \Delta_{\max} = \mathcal{O}(\eta^b).
\end{aligned}
\tag{14}
$$

534 Since $\eta < (2L)^{-1}$, we have that $\eta L < 1$ and for any $\epsilon > 0$, we can solve for b such that
535 $\eta^b L^{(b-1)} \Delta_{\max} < \epsilon$. Therefore the sequence $\{\boldsymbol{\omega}_t^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Moreover, in a
536 complete space, every Cauchy sequence has a limit: $\lim_{k\to\infty} \boldsymbol{\omega}_t^{(k)} = \boldsymbol{\omega}_t^*$ $\qquad\square$

A.2 Proof of Theorem 5.1

538 **Theorem A.1.** *Consider the (Minimax) problem under Assumption 3.1 and Lv.k GP. Let $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$*
539 *be a stationary point. Suppose $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ not in kernel of $\nabla_{\phi\theta} f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$, $\boldsymbol{\phi}_t - \boldsymbol{\phi}^*$ not in kernel of*
540 *$\nabla_{\theta\phi} f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ and $\eta < (L)^{-1}$. There exists a neighborhood $\mathcal{U}$ of $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ such that if SPPM started*
541 *at $(\boldsymbol{\theta}_0, \boldsymbol{\phi}_0) \in \mathcal{U}$, the iterates $\{\boldsymbol{\theta}_t, \boldsymbol{\phi}_t\}_{t\geq 0}$ generated by SPPM satisfy:*

$$
\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2 \leq \frac{\rho^2(\boldsymbol{I} - \eta\nabla_{\theta\theta} f^*)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \rho^2(\boldsymbol{I} + \eta\nabla_{\phi\phi} f^*)\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{1 + \eta^2 \lambda_{\min}(\nabla_{\theta\phi} f^* \nabla_{\phi\theta} f^*)}
$$

542 *where $f^* = f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$. Moreover, for any η satisfying:*

$$
\frac{\max(\rho^2(\boldsymbol{I} - \eta\nabla_{\theta\theta} f^*), \rho^2(\boldsymbol{I} + \eta\nabla_{\phi\phi} f^*))}{1 + \eta^2 \lambda_{\min}(\nabla_{\theta\phi} f^* \nabla_{\phi\theta} f^*)} < 1,
\tag{15}
$$

543 *SPPM converges asymptotically to $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$.*

544 *Proof.* Consider the learning dynamics:

$$
\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_{t+1}) \\
\boldsymbol{\phi}_{t+1} &= \boldsymbol{\phi}_t + \eta\nabla_\theta f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_t)
\end{aligned}
$$

545 Let us define

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_t &= \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \\
\hat{\boldsymbol{\phi}}_t &= \boldsymbol{\phi}_t - \boldsymbol{\phi}^*
\end{aligned}
$$

546 It follows immediately by linearizing the system about the stationary point $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ that

$$
\begin{bmatrix} \hat{\boldsymbol{\theta}}_{t+1} \\ \hat{\boldsymbol{\phi}}_{t+1} \end{bmatrix} \simeq \begin{bmatrix} \boldsymbol{I} - \eta\nabla_{\theta\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} + \eta\nabla_{\phi\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_t \\ \hat{\boldsymbol{\phi}}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & -\eta\nabla_{\theta\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \\ \eta\nabla_{\phi\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_{t+1} \\ \hat{\boldsymbol{\phi}}_{t+1} \end{bmatrix}
$$

547 Let us denote the Jacobian by

$$
\begin{bmatrix} -\nabla_{\theta\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & -\nabla_{\theta\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \\ \nabla_{\phi\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & \nabla_{\phi\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \end{bmatrix} = \begin{bmatrix} -\boldsymbol{A} & -\boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{bmatrix}
\tag{16}
$$

16

Then we can rewrite the dynamics around the stationary point as

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{A}\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B}\hat{\boldsymbol{\phi}}_{t+1}$$

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{A}\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B}(\hat{\boldsymbol{\phi}}_t + \eta \boldsymbol{B}^T \hat{\boldsymbol{\theta}}_{t+1} + \eta \boldsymbol{C}\hat{\boldsymbol{\phi}}_t)$$

$$(\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)\hat{\boldsymbol{\theta}}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B}(\boldsymbol{I} + \eta \boldsymbol{C})\boldsymbol{\phi}_t$$

$$\hat{\boldsymbol{\theta}}_{t+1} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}\left[(\boldsymbol{I} - \eta \boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B}(\boldsymbol{I} + \eta \boldsymbol{C})\boldsymbol{\phi}_t\right] \quad (17)$$

Similarly, for the other player we have

$$\hat{\boldsymbol{\phi}}_{t+1} = \hat{\boldsymbol{\phi}}_t + \eta \boldsymbol{B}^T \hat{\boldsymbol{\theta}}_{t+1} + \eta \boldsymbol{C}\hat{\boldsymbol{\phi}}_t$$

$$\hat{\boldsymbol{\phi}}_{t+1} = \hat{\boldsymbol{\phi}}_t + \eta \boldsymbol{B}^T (\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{A}\hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B}\hat{\boldsymbol{\phi}}_{t+1}) + \eta \boldsymbol{C}\hat{\boldsymbol{\phi}}_t$$

$$(\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})\hat{\boldsymbol{\phi}}_{t+1} = \eta \boldsymbol{B}^T (\boldsymbol{I} - \eta \boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta \boldsymbol{C})\boldsymbol{\phi}_t$$

$$\hat{\boldsymbol{\phi}}_{t+1} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1}\left[\eta \boldsymbol{B}^T (\boldsymbol{I} - \eta \boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta \boldsymbol{C})\boldsymbol{\phi}_t\right] \quad (18)$$

Let us define the symmetric matrices $\boldsymbol{Q}_{\boldsymbol{\theta}} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}$, $\boldsymbol{Q}_{\boldsymbol{\phi}} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1}$ and $\boldsymbol{P}_{\boldsymbol{\theta}} = (\boldsymbol{I} - \eta \boldsymbol{A})$, $\boldsymbol{P}_{\boldsymbol{\phi}} = (\boldsymbol{I} + \eta \boldsymbol{C})$. Further we define $r_t = \|\hat{\boldsymbol{\theta}}_{t+1}\|^2 + \|\hat{\boldsymbol{\phi}}_{t+1}\|^2$. Based on these definitions, and the expressions in (53) and (54) we have

$$\|\boldsymbol{\theta}_{t+1}\|^2 + \|\boldsymbol{\phi}_{t+1}\|^2 = \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 + \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2$$
$$- 2\eta \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t + 2\eta \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$(19)$$

To simplify the expression in (55) we use the following lemma:

Lemma A.1. *The matrices* $\boldsymbol{Q}_{\boldsymbol{\theta}} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}$, $\boldsymbol{Q}_{\boldsymbol{\phi}} = (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1}$ *satisfy the following properties:*

$$\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{\phi}} \quad (20)$$

$$\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T = \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \quad (21)$$

Using this lemma, we can show that

$$\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t = \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^T \boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t = \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$

where the intermediate equality holds as $\boldsymbol{a}^T \boldsymbol{b} = \boldsymbol{b}^T \boldsymbol{a}$. Hence, the expression in (55) can be simplified as

$$\|\hat{\boldsymbol{\theta}}_{t+1}\|^2 + \|\hat{\boldsymbol{\phi}}_{t+1}\|^2 = \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 + \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 \quad (22)$$

We simplify equation (58) as follows. Consider the term involving $\hat{\boldsymbol{\theta}}_t$. We have

$$\|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 = \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t + \eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$= \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{Q}_{\boldsymbol{\theta}}^2 + \eta^2 \boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{B}^T)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$= \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{Q}_{\boldsymbol{\theta}}^2 + \eta^2 \boldsymbol{B}\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}})\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$= \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{Q}_{\boldsymbol{\theta}}^2 + \eta^2 \boldsymbol{B}\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{Q}_{\boldsymbol{\theta}})\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$= \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)\boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t$$
$$= \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t \quad (23)$$

where the last equality follows by replacing $\boldsymbol{Q}_{\boldsymbol{\theta}}$ by its definition. The same procedure follows for the term involving $\hat{\boldsymbol{\phi}}_t$ which leads to the expression

$$\|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 = \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t. \quad (24)$$

Substitute $\|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t\|^2$ and $\|\boldsymbol{Q}_{\boldsymbol{\phi}}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2 + \eta^2 \|\boldsymbol{Q}_{\boldsymbol{\theta}}\boldsymbol{B}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t\|^2$ in (58) with the expressions in (23) and (24), respectively, to obtain

$$\|\hat{\boldsymbol{\theta}}_{t+1}\|^2 + \|\hat{\boldsymbol{\phi}}_{t+1}\|^2 = \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t + \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1}\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t. \quad (25)$$

17

564 Note that, we assume that the trajectory $\{\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\phi}}_t\}_{t \geq 0}$ is not in the kernel of $\boldsymbol{BB}^T$ and $\boldsymbol{B}^T\boldsymbol{B}$, thus
565 $\boldsymbol{BB}^T\hat{\boldsymbol{\theta}}_t \neq 0$ and $\boldsymbol{B}^T\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \neq 0$. Now using the expression in (61) and the fact that $\boldsymbol{P_\theta} = \boldsymbol{P_\theta^T}$,
566 $\boldsymbol{P_\phi} = \boldsymbol{P_\phi^T}$ and $\boldsymbol{BB}^T$ and $\boldsymbol{B}^T\boldsymbol{B}$ have the same set of non-zero eigenvalues, if we denote the minimum
567 non-zero eigenvalues by $\lambda_{\min}(\boldsymbol{BB}^T)$ and $\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})$, we can write

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2 \leq \frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \rho^2(1 + \eta\boldsymbol{C})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})}.$$

568 Replacing $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2$ and $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2$ with r_{t+1} and r_t we have:

$$r_{t+1} \leq \frac{\max(\rho^2(\boldsymbol{I} - \eta\boldsymbol{A}), \rho^2(\boldsymbol{I} + \eta\boldsymbol{C}))}{1 + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})}r_t.$$

569 Recall that $\boldsymbol{A} = \nabla_{\theta\theta}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$, $\boldsymbol{B} = \nabla_{\theta\phi}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ and $\boldsymbol{C} = \nabla_{\phi\phi}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$, therefore for any η
570 satisfying that:

$$\frac{\max(\rho^2(\boldsymbol{I} - \eta\nabla_{\theta\theta}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)), \rho^2(\boldsymbol{I} + \eta\nabla_{\phi\phi}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)))}{1 + \eta^2\lambda_{\min}(\nabla_{\theta\phi}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)\nabla_{\phi\theta}f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*))} < 1, \tag{26}$$

571 we have $r_{t+1} < r_t$. Since we linearize the system about the stationary point $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$, there exists
572 a neighborhood $\mathcal{U}$ around the stationary point, such that, SPPM started at $(\boldsymbol{\theta}_0, \boldsymbol{\phi}_0) \in \mathcal{U}$ converges
573 asymptotically to $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$. $\qquad\square$

A.3 Proof of Remark 5.1

575 *Proof.* To prove the local convergence of Lv.k GP in non-convex non-concave games, we first
576 consider the update rule of Lv.k GP:

$$\text{Reasoning:} \begin{cases} \boldsymbol{\theta}_t^{(k)} = \boldsymbol{\theta}_t - \eta\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(k-1)}) \\ \boldsymbol{\phi}_t^{(k)} = \boldsymbol{\phi}_t - \eta\nabla_{\boldsymbol{\phi}}g(\boldsymbol{\theta}_t^{(k-1)}, \boldsymbol{\phi}_t) \end{cases} \quad \text{Update:} \begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t^{(k)} \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t^{(k)} \end{cases}$$

577 Similar to Section A.2 let us denote

$$\begin{bmatrix} -\nabla_{\theta\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & -\nabla_{\theta\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \\ \nabla_{\phi\theta}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) & \nabla_{\phi\phi}^2 f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \end{bmatrix} = \begin{bmatrix} -\boldsymbol{A} & -\boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{bmatrix}$$

578 and we define the difference between states and stationary points as

$$\hat{\boldsymbol{\theta}}_t^{(k)} = \boldsymbol{\theta}_t^{(k)} - \boldsymbol{\theta}^* \text{ and } \hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^*$$
$$\hat{\boldsymbol{\phi}}_t^{(k)} = \boldsymbol{\phi}_t^{(k)} - \boldsymbol{\phi}^* \text{ and } \hat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_t - \boldsymbol{\phi}^*$$

579 Linearizing the dynamical system induced by Lv.k GP about the stationary point $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ we get:

$$\begin{cases} \hat{\boldsymbol{\theta}}_{t+1} & = \hat{\boldsymbol{\theta}}_t^{(k)} = (\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\hat{\boldsymbol{\phi}}_t^{(k-1)} \\ \hat{\boldsymbol{\phi}}_{t+1} & = \hat{\boldsymbol{\phi}}_t^{(k)} = \eta\boldsymbol{B}^T\hat{\boldsymbol{\theta}}_t^{(k-1)} + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t \end{cases}$$

580 Note, in Lv.k GP, we define $\boldsymbol{\theta}_t^{(0)} = \boldsymbol{\theta}_t$ and $\boldsymbol{\phi}_t^{(0)} = \boldsymbol{\phi}_t$, thus for Lv.1 GP, we have:

$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(1)} = (\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \\ \hat{\boldsymbol{\phi}}_t^{(1)} = \eta\boldsymbol{B}^T\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t \end{cases}$$

581 For Lv.2 GP, we have:

$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(2)} = (\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\hat{\boldsymbol{\phi}}_t^{(1)} \\ \hat{\boldsymbol{\phi}}_t^{(2)} = \eta\boldsymbol{B}^T\hat{\boldsymbol{\theta}}_t^{(1)} + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t \end{cases}$$

582 Substituting $\hat{\boldsymbol{\theta}}_t^{(1)}$ and $\hat{\boldsymbol{\phi}}_t^{(1)}$ into the update rule above we get:

$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(2)} = (\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}(\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t - \eta^2\boldsymbol{BB}^T\hat{\boldsymbol{\theta}}_t \\ \hat{\boldsymbol{\phi}}_t^{(2)} = \eta\boldsymbol{B}^T(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t - \eta^2\boldsymbol{B}^T\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \end{cases}$$

18

Similarly, for Lv.3 and Lv.4 GP we have:
$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(3)} = (\boldsymbol{I} - \eta^2\boldsymbol{BB}^T)(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}(\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t + \eta^3\boldsymbol{BB}^T\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \\ \hat{\boldsymbol{\phi}}_t^{(3)} = \eta\boldsymbol{B}^T(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} - \eta^2\boldsymbol{B}^T\boldsymbol{B})(\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t - \eta^3\boldsymbol{BB}^T\boldsymbol{B}\hat{\boldsymbol{\theta}}_t \end{cases}$$

and
$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(4)} = (\boldsymbol{I} - \eta^2\boldsymbol{BB}^T)\left[(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}(\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t\right] + \eta^4\boldsymbol{BB}^T\boldsymbol{BB}^T\hat{\boldsymbol{\theta}}_t \\ \hat{\boldsymbol{\phi}}_t^{(4)} = (\boldsymbol{I} - \eta^2\boldsymbol{B}^T\boldsymbol{B})\left[\eta\boldsymbol{B}^T(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t\right] + \eta^4\boldsymbol{B}^T\boldsymbol{BB}^T\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \end{cases}$$

Summarizing the equations above we have that for Lv.$2k$ GP, its update can be written as:
$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(2k)} = (\sum_{i=0}^{k-1}(-\eta^2\boldsymbol{BB}^T)^k)\left[(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}(\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t\right] + (-\eta^2\boldsymbol{BB}^T)^k\hat{\boldsymbol{\theta}}_t \\ \hat{\boldsymbol{\phi}}_t^{(2k)} = (\sum_{i=0}^{k-1}(-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k)\left[\eta\boldsymbol{B}^T(\boldsymbol{I} - \eta\boldsymbol{A})\hat{\boldsymbol{\theta}}_t + (\boldsymbol{I} + \eta\boldsymbol{C})\hat{\boldsymbol{\phi}}_t\right] + (-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k\hat{\boldsymbol{\phi}}_t \end{cases}$$

Similar to Appendix A.2, let us define $\boldsymbol{Q_\theta} = (\boldsymbol{I} + \eta^2\boldsymbol{BB}^T)^{-1}$, $\boldsymbol{Q_\phi} = (\boldsymbol{I} + \eta^2\boldsymbol{B}^T\boldsymbol{B})^{-1}$ and $\boldsymbol{P_\theta} = (\boldsymbol{I} - \eta\boldsymbol{A})$, $\boldsymbol{P_\phi} = (\boldsymbol{I} + \eta\boldsymbol{C})$. Further, we define $\boldsymbol{R_\theta}^{(k)} = (\sum_{i=0}^{k-1}(-\eta^2\boldsymbol{BB}^T)^k)$, $\boldsymbol{R_\phi}^{(k)} = (\sum_{i=0}^{k-1}(-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k)$ and $\boldsymbol{E_\theta}^{(k)} = \boldsymbol{R_\theta}^{(k)} - \boldsymbol{Q_\theta}$, $\boldsymbol{E_\phi}^{(k)} = \boldsymbol{R_\phi}^{(k)} - \boldsymbol{Q_\phi}$.

Since $\eta < L^{-1}$, we have that:
$$\boldsymbol{Q_\theta} = \sum_{i=0}^{\infty}(-\eta^2\boldsymbol{BB}^T)^k$$
$$\boldsymbol{Q_\phi} = \sum_{i=0}^{\infty}(-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k$$

and
$$\boldsymbol{E_\theta}^{(k)} = \boldsymbol{R_\theta}^{(k)} - \boldsymbol{Q_\theta} = -\sum_{i=k}^{\infty}(-\eta^2\boldsymbol{BB}^T)^k = -(\boldsymbol{I} + \eta^2\boldsymbol{BB}^T)^{-1} \cdot (-\eta^2\boldsymbol{BB}^T)^k \qquad (27)$$

$$\boldsymbol{E_\phi}^{(k)} = \boldsymbol{R_\phi}^{(k)} - \boldsymbol{Q_\phi} = -\sum_{i=k}^{\infty}(-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k = -(\boldsymbol{I} + \eta^2\boldsymbol{B}^T\boldsymbol{B})^{-1} \cdot (-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k \qquad (28)$$

Also, from Lemma A.1 and the definition of the error terms, it can be verified that
$$\boldsymbol{E_\theta}^{(k)}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{E_\phi}^{(k)} \qquad (29)$$
$$\boldsymbol{E_\phi}^{(k)}\boldsymbol{B}^T = \boldsymbol{B}^T\boldsymbol{E_\theta}^{(k)} \qquad (30)$$

Then we can rewrite the update rule of Lv.$2k$ GP:
$$\begin{cases} \hat{\boldsymbol{\theta}}_t^{(2k)} = (\boldsymbol{Q_\theta} + \boldsymbol{E_\theta}^{(k)})\left[\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right] + (-\eta^2\boldsymbol{BB}^T)^k\hat{\boldsymbol{\theta}}_t \\ \hat{\boldsymbol{\phi}}_t^{(2k)} = (\boldsymbol{Q_\phi} + \boldsymbol{E_\phi}^{(k)})\left[\eta\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t + \boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right] + (-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k\hat{\boldsymbol{\phi}}_t \end{cases}$$

Let us consider the following sum:
$$\|\hat{\boldsymbol{\theta}}_t^{(2k)} - (-\eta^2\boldsymbol{BB}^T)^k\hat{\boldsymbol{\theta}}_t\|^2 + \|\hat{\boldsymbol{\phi}}_t^{(2k)} - (-\eta^2\boldsymbol{B}^T\boldsymbol{B})^k\hat{\boldsymbol{\phi}}_t\|^2$$
$$= \left\|(\boldsymbol{Q_\theta} + \boldsymbol{E_\theta}^{(k)})\left[\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right]\right\|^2 + \left\|(\boldsymbol{Q_\phi} + \boldsymbol{E_\phi}^{(k)})\left[\eta\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t + \boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right]\right\|^2 \qquad (31)$$

The R.H.S. of Eq.(31) can be written as:
$$\left\|(\boldsymbol{Q_\theta} + \boldsymbol{E_\theta}^{(k)})\left[\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - \eta\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right]\right\|^2 + \left\|(\boldsymbol{Q_\phi} + \boldsymbol{E_\phi}^{(k)})\left[\eta\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t + \boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t\right]\right\|^2$$
$$= \hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{Q_\theta}^2\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - 2\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{Q_\theta}^2\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + \eta^2\hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T\boldsymbol{B}^T\boldsymbol{Q_\theta}^2\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t$$
$$+ 2\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{Q_\theta}\boldsymbol{E_\theta}^{(k)}\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - 4\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{Q_\theta}\boldsymbol{E_\theta}^{(k)}\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + 2\eta^2\hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T\boldsymbol{B}^T\boldsymbol{Q_\theta}\boldsymbol{E_\theta}^{(k)}\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t$$
$$+ \hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T[\boldsymbol{E_\theta}^{(k)}]^2\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t - 2\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T[\boldsymbol{E_\theta}^{(k)}]^2\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + \eta^2\hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T\boldsymbol{B}^T[\boldsymbol{E_\theta}^{(k)}]^2\boldsymbol{B}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t$$
$$+ \hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T\boldsymbol{Q_\phi}^2\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + 2\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}\boldsymbol{Q_\phi}^2\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + \eta^2\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}\boldsymbol{Q_\phi}^2\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t$$
$$+ 2\hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T\boldsymbol{Q_\phi}\boldsymbol{E_\phi}^{(k)}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + 4\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}\boldsymbol{Q_\phi}\boldsymbol{E_\phi}^{(k)}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + 2\eta^2\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}\boldsymbol{Q_\phi}\boldsymbol{E_\phi}^{(k)}\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t$$
$$+ \hat{\boldsymbol{\phi}}_t^T\boldsymbol{P_\phi}^T[\boldsymbol{E_\phi}^{(k)}]^2\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + 2\eta\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}[\boldsymbol{E_\phi}^{(k)}]^2\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t + \eta^2\hat{\boldsymbol{\theta}}_t^T\boldsymbol{P_\theta}^T\boldsymbol{B}[\boldsymbol{E_\phi}^{(k)}]^2\boldsymbol{B}^T\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t \qquad (32)$$

19

595 Now, before adding all terms in Eq.(32), note that all of the cross terms in Eq.(32) cancel out.

596 For instance, using Lemma A.1 and Eq.(29), Eq.(30) we can show that

$$
\begin{aligned}
4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t - 4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t &= 4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{B} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t - 4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \\
&= 4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t - 4\eta\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \\
&= 0
\end{aligned}
$$

597 By using similar arguments it can be shown that terms in Eq.(32) leads to:

$$
\begin{aligned}
&\left\| (\boldsymbol{Q}_{\boldsymbol{\theta}} + \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}) \left[\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 + \left\| (\boldsymbol{Q}_{\boldsymbol{\phi}} + \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}) \left[\eta \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 \\
=&\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \eta^2 \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \\
&+ 2\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + 2\eta^2 \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \\
&+ \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \eta^2 \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \\
&+ \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t + \eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} \boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t \\
&+ 2\hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t + 2\eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t \\
&+ \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t + \eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t
\end{aligned}
\tag{33}
$$

598 Similar to Eq.(23) we have the following simplification:

$$
\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} \boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t
$$

$$
\hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}}^2 \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t + \eta^2 \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}}^2 \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t = \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t
$$

$$
2\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + 2\eta^2 \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{B} \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t = 2\hat{\boldsymbol{\theta}}_t^T \boldsymbol{P}_{\boldsymbol{\theta}}^T \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t
$$

$$
2\hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{Q}_{\boldsymbol{\phi}} \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t + 2\eta^2 \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\theta}} \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t = 2\hat{\boldsymbol{\phi}}_t^T \boldsymbol{P}_{\boldsymbol{\phi}}^T \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t
$$

599 Now we can further simplify Eq.(32) as:

$$
\begin{aligned}
&\left\| (\boldsymbol{Q}_{\boldsymbol{\theta}} + \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}) \left[\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 + \left\| (\boldsymbol{Q}_{\boldsymbol{\phi}} + \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}) \left[\eta \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 \\
=&(\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T \left[\boldsymbol{Q}_{\boldsymbol{\theta}} + 2\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} + [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 + \eta^2 \boldsymbol{B} [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 \boldsymbol{B}^T \right] (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
&+ (\boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)^T \left[\boldsymbol{Q}_{\boldsymbol{\phi}} + 2\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} + [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 + \eta^2 \boldsymbol{B} [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 \boldsymbol{B}^T \right] (\boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)
\end{aligned}
$$

600 Using Eq.(27) and Eq.(28) and definition of $\boldsymbol{Q}_{\boldsymbol{\theta}}$ and $\boldsymbol{Q}_{\boldsymbol{\phi}}$ we have:

$$
\begin{aligned}
&(\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T \left[\boldsymbol{Q}_{\boldsymbol{\theta}} + 2\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)} + [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 + \eta^2 \boldsymbol{B} [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 \boldsymbol{B}^T \right] (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
=&(\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-1} (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) - 2(\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-1} (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^k (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
&+ (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T) (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-2} (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{2k} (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
=&(\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-1} (\boldsymbol{I} - 2(-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{(k)} + (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{2k}) (\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
=&((\boldsymbol{I} - (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^k) \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-1} ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^k) \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)
\end{aligned}
$$

601 Similarly, we have that

$$
\begin{aligned}
&(\boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)^T \left[\boldsymbol{Q}_{\boldsymbol{\phi}} + 2\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)} + [\boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}]^2 + \eta^2 \boldsymbol{B}^T [\boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}]^2 \boldsymbol{B} \right] (\boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t) \\
=&((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T \boldsymbol{B})^k) \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1} ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T \boldsymbol{B})^k) \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)
\end{aligned}
$$

602 Thus we simplify the R.H.S. of Eq.(31) as

$$
\begin{aligned}
&\left\| (\boldsymbol{Q}_{\boldsymbol{\theta}} + \boldsymbol{E}_{\boldsymbol{\theta}}^{(k)}) \left[\boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t - \eta \boldsymbol{B} \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 + \left\| (\boldsymbol{Q}_{\boldsymbol{\phi}} + \boldsymbol{E}_{\boldsymbol{\phi}}^{(k)}) \left[\eta \boldsymbol{B}^T \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t + \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t \right] \right\|^2 \\
=&((\boldsymbol{I} - (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^k) \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B} \boldsymbol{B}^T)^{-1} ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B} \boldsymbol{B}^T)^k) \boldsymbol{P}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}_t) \\
&+ ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T \boldsymbol{B})^k) \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T \boldsymbol{B})^{-1} ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T \boldsymbol{B})^k) \boldsymbol{P}_{\boldsymbol{\phi}} \hat{\boldsymbol{\phi}}_t)
\end{aligned}
\tag{34}
$$

20

Let us consider the L.H.S. of Eq.(31)

$$
\|\hat{\boldsymbol{\theta}}_t^{(2k)} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t\|^2 + \|\hat{\boldsymbol{\phi}}_t^{(2k)} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t\|^2
$$
$$
= \|\hat{\boldsymbol{\theta}}_t^{(2k)}\|^2 - 2\langle \hat{\boldsymbol{\theta}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t \rangle + \|(-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t\|^2
$$
$$
+ \|\hat{\boldsymbol{\phi}}_t^{(2k)}\|^2 - 2\langle \hat{\boldsymbol{\phi}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t \rangle + \|(-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t\|^2 \tag{35}
$$

Substituting Eq.(35) and Eq.(34) into L.H.S. and R.H.S. of Eq.(31) respectively we get:

$$
\|\hat{\boldsymbol{\theta}}_t^{(2k)}\|^2 - 2\langle \hat{\boldsymbol{\theta}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t \rangle + \|(-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t\|^2
$$
$$
+ \|\hat{\boldsymbol{\phi}}_t^{(2k)}\|^2 - 2\langle \hat{\boldsymbol{\phi}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t \rangle + \|(-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t\|^2
$$
$$
= ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)
$$
$$
+ ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T\boldsymbol{B})^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)
$$

Now we have the following equation:

$$
\|\hat{\boldsymbol{\theta}}_t^{(2k)}\|^2 + \|\hat{\boldsymbol{\phi}}_t^{(2k)}\|^2
$$
$$
= ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)
$$
$$
+ 2\langle \hat{\boldsymbol{\theta}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t \rangle - \|(-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t\|^2
$$
$$
+ ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T\boldsymbol{B})^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)
$$
$$
+ 2\langle \hat{\boldsymbol{\phi}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t \rangle - \|(-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t\|^2
$$

Note that

$$
2\langle \hat{\boldsymbol{\theta}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t \rangle = 2\langle (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{\frac{k}{2}} \hat{\boldsymbol{\theta}}_t^{(2k)}, (\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{\frac{k}{2}} \hat{\boldsymbol{\theta}}_t \rangle \tag{36}
$$
$$
\leq \eta^{2k} \hat{\boldsymbol{\theta}}_t^{(2k)T} (\boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t^{(2k)} + \eta^{2k} \hat{\boldsymbol{\theta}}_t^T (\boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t \tag{37}
$$

Similarly

$$
2\langle \hat{\boldsymbol{\phi}}_t^{(2k)}, (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\phi}}_t \rangle = 2\langle (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^{\frac{k}{2}} \hat{\boldsymbol{\phi}}_t^{(2k)}, (\eta^2 \boldsymbol{B}^T\boldsymbol{B})^{\frac{k}{2}} \hat{\boldsymbol{\phi}}_t \rangle \tag{38}
$$
$$
\leq \eta^{2k} \hat{\boldsymbol{\phi}}_t^{(2k)T} (\boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t^{(2k)} + \eta^{2k} \hat{\boldsymbol{\phi}}_t^T (\boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t \tag{39}
$$

Summing everything together we have:

$$
(\hat{\boldsymbol{\theta}}_t^{(2k)})^T (\boldsymbol{I} - (\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)(\hat{\boldsymbol{\theta}}_t^{(2k)}) + (\hat{\boldsymbol{\phi}}_t^{(2k)})^T (\boldsymbol{I} - (\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)(\hat{\boldsymbol{\phi}}_t^{(2k)})
$$
$$
\leq ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}\boldsymbol{B}^T)^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k)\boldsymbol{P}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}_t)
$$
$$
+ (\hat{\boldsymbol{\theta}}_t)^T (\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k (\hat{\boldsymbol{\theta}}_t) - \|(-\eta^2 \boldsymbol{B}\boldsymbol{B}^T)^k \hat{\boldsymbol{\theta}}_t\|^2
$$
$$
+ ((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)^T (\boldsymbol{I} + \eta^2 \boldsymbol{B}^T\boldsymbol{B})^{-1}((\boldsymbol{I} - (-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k)\boldsymbol{P}_{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}_t)
$$
$$
+ (\hat{\boldsymbol{\phi}}_t)^T (\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k (\hat{\boldsymbol{\phi}}_t) - \|(-\eta^2 \boldsymbol{B}^T\boldsymbol{B})^k \hat{\boldsymbol{\phi}}_t\|^2
$$

Note that, we assume that the trajectory $\{\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\phi}}_t\}_{t \geq 0}$ is not in the kernel of $\boldsymbol{B}\boldsymbol{B}^T$ and $\boldsymbol{B}^T\boldsymbol{B}$, thus $\boldsymbol{B}\boldsymbol{B}^T\hat{\boldsymbol{\theta}}_t \neq 0$ and $\boldsymbol{B}^T\boldsymbol{B}\hat{\boldsymbol{\phi}}_t \neq 0$. Now using the expression in (61) and the fact that $\boldsymbol{P}_{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{\theta}}^T$, $\boldsymbol{P}_{\boldsymbol{\phi}} = \boldsymbol{P}_{\boldsymbol{\phi}}^T$ and $\boldsymbol{B}\boldsymbol{B}^T$ and $\boldsymbol{B}^T\boldsymbol{B}$ have the same set of non-zero eigenvalues, if we denote the minimum

21

612 non-zero eigenvalues by $\lambda_{\min}(BB^T)$ and $\lambda_{\min}(B^TB)$, we can write

$$(1 - (\eta^2\lambda_{max}(BB^T))^k)\|\theta_t^{(\hat{2}k)}\|^2 + (1 - (\eta^2\lambda_{max}(B^TB))^k)\|\phi_t^{(\hat{2}k)}\|^2$$

$$\leq \frac{(1 - (-\eta^2\lambda(BB^T))^k)^2\rho^2(1-\eta A)\|\hat{\theta}_t\|^2}{1 + \eta^2\lambda_{min}(BB^T)} + (\eta^2\lambda(BB^T))^k\|\hat{\theta}_t\|^2 - (\eta^2\lambda(BB^T))^{2k}\|\hat{\theta}_t\|^2$$

$$+ \frac{(1 - (-\eta^2\lambda(B^TB))^k)^2\rho^2(1+\eta C)\|\hat{\phi}_t\|^2}{1 + \eta^2\lambda_{min}(B^TB)} + (\eta^2\lambda(B^TB))^k\|\hat{\phi}_t\|^2 - (\eta^2\lambda(B^TB))^{2k}\|\hat{\phi}_t\|^2$$

$$\leq \frac{(1 - (-\eta^2\lambda(BB^T))^k)^2\rho^2(1-\eta A)\|\hat{\theta}_t\|^2}{1 + \eta^2\lambda_{min}(BB^T)} + (\eta^2\lambda(BB^T))^k(1 - (\eta^2\lambda(BB^T))^k)\|\hat{\theta}_t\|^2$$

$$+ \frac{(1 - (-\eta^2\lambda(B^TB))^k)^2\rho^2(1+\eta C)\|\hat{\phi}_t\|^2}{1 + \eta^2\lambda_{min}(B^TB)} + (\eta^2\lambda(B^TB))^k\|\hat{\phi}_t\|^2 - (\eta^2\lambda(B^TB))^{2k}\|\hat{\phi}_t\|^2$$

$$\leq \frac{(1 - (-\eta^2\lambda(BB^T))^k)^2\rho^2(1-\eta A)\|\hat{\theta}_t\|^2}{(1 + \eta^2\lambda_{min}(BB^T))(1 - (\eta^2\lambda_{max}(BB^T))^k)} + \frac{(\eta^2\lambda(BB^T))^k(1 - (\eta^2\lambda(BB^T))^k)\|\hat{\theta}_t\|^2}{(1 - (\eta^2\lambda_{max}(BB^T))^k)}$$

$$+ \frac{(1 - (-\eta^2\lambda(B^TB))^k)^2\rho^2(1+\eta C)\|\hat{\phi}_t\|^2}{(1 + \eta^2\lambda_{min}(B^TB))(1 - (\eta^2\lambda_{max}(B^TB))^k)} + \frac{(\eta^2\lambda(B^TB))^k(1 - (\eta^2\lambda(B^TB))^k)\|\hat{\phi}_t\|^2}{(1 - (\eta^2\lambda_{max}(B^TB))^k)}$$

(40)

613 Let us define the distance as:

$$\|r_t^{(k)}\|^2 = \|\hat{\theta}_t^{(k)}\|^2 + \|\hat{\phi}_t^{(k)}\|^2 \tag{41}$$

$$\|r_t\|^2 = \|\hat{\theta}_t\|^2 + \|\hat{\phi}_t\|^2 \tag{42}$$

614 Then we have

$$\|r_t^{(2k)}\| \leq \frac{(1 - (-\eta^2\lambda(BB^T))^k)(\rho^2(1-\eta A)\|\hat{\theta}_t\|^2 + \rho^2(1+\eta C)\|\hat{\phi}_t\|^2)}{(1 + \eta^2\lambda_{min}(BB^T))(1 - (\eta^2\lambda_{max}(BB^T))^k)} + \frac{(\eta^2\lambda(BB^T))^k(1 - (\eta^2\lambda(BB^T))^k)\|r_t\|^2}{(1 - (\eta^2\lambda_{max}(BB^T))^k)}$$

$$= a\left(\frac{\rho^2(1-\eta A)\|\hat{\theta}_t\|^2 + \rho^2(1+\eta C)\|\hat{\phi}_t\|^2}{1 + \eta^2\lambda_{min}(BB^T)}\right) + b \tag{43}$$

615 where

$$a = \begin{cases} \frac{1 + (\eta^2\lambda_{\max}(BB^T))^k}{1 - (\eta^2\lambda_{\max}(BB^T))^k} \text{ for odd } k, \\ \frac{1 - (\eta^2\lambda_{\min}(BB^T))^k}{1 - (\eta^2\lambda_{\max}(BB^T))^k} \text{ for even } k \end{cases}$$

616 and

$$b = \frac{(\eta^2\lambda_{\max}(BB^T))^k(1 - (\eta^2\lambda_{\min}(BB^T))^k)}{1 - (\eta^2\lambda_{\max}(BB^T))^k} \tag{44}$$

617 $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

618 A.3.1 Proof of Lemma A.1

619 Let $B = U\Sigma V^T$ be the singular value decomposition of B. Here U and V are orthonormal matrices
620 and Σ is a rectangular diagonal matrix. Then we have:

$$Q_\theta B = (I + \eta^2 U\Sigma V^T V\Sigma^T U^T)^{-1} U\Sigma V^T$$
$$= (U(\eta^2\Sigma\Sigma^T + I)U^T)^{-1} U\Sigma V^T$$
$$= U(\eta^2\Sigma\Sigma^T + I)^{-1} U^T U\Sigma V^T$$
$$= U(\eta^2\Sigma\Sigma^T + I)^{-1}\Sigma V^T$$

621 Here we used the fact that U and V are orthonormal matrices. Now, we simplify the other side to
622 get:

$$BQ_\phi = U\Sigma V^T(I + \eta^2 V\Sigma^T U^T U\Sigma V^T)^{-1}$$
$$= U\Sigma V^T(V(\eta^2\Sigma^T\Sigma + I)V^T)^{-1}$$
$$= U\Sigma V^T V(\eta^2\Sigma^T\Sigma + I)^{-1}V^T$$
$$= U\Sigma(\eta^2\Sigma^T\Sigma + I)^{-1}V^T$$

22

623 Now we consider the following equation:

$$\eta^2 \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I}) = (\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})\boldsymbol{\Sigma} \tag{45}$$

624 which indicates that $\boldsymbol{\Sigma}(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I}) = (\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})\boldsymbol{\Sigma}$. Multiplying both sides of this equation by
625 $(\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})^{-1}$ and $(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I})^{-1}$ we have:

$$(\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I})(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I})^{-1} = (\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})^{-1}(\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})\boldsymbol{\Sigma}(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I})^{-1}$$
$$(\eta^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T + \boldsymbol{I})^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\eta^2\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \boldsymbol{I})^{-1}$$

626 Therefore, we have $\boldsymbol{Q_\theta}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{Q_\phi}$. Using a similar argument, we can also prove the equality in
627 Equation (21).

A.4 Theorem 5.1 without kernel assumption

629 **Theorem A.2.** *Consider the (Minimax) problem under Assumption 3.1 and Lv.k GP. Let $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$*
630 *be a stationary point. Suppose $\eta < (L)^{-1}$. There exists a neighborhood $\mathcal{U}$ of $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ such that if*
631 *SPPM started at $(\boldsymbol{\theta}_0, \boldsymbol{\phi}_0) \in \mathcal{U}$, the iterates $\{\boldsymbol{\theta}_t, \boldsymbol{\phi}_t\}_{t \geq 0}$ generated by SPPM satisfy:*

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2 \leq \frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}\boldsymbol{B}^T)} + \frac{\rho^2(1 + \eta\boldsymbol{C})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})}.$$

632 *where $f^* = f(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$. Moreover, for any η satisfying:*

$$\frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}\boldsymbol{B}^T)} + \frac{\rho^2(1 + \eta\boldsymbol{C})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})} < \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi_t} - \boldsymbol{\phi^*}\|^2 \tag{46}$$

633 *SPPM converges asymptotically to $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$.*

634 *Proof.* Following the same setting and procedure as in Appendix A.2, we have that

$$\|\hat{\boldsymbol{\theta}}_{t+1}\|^2 + \|\hat{\boldsymbol{\phi}}_{t+1}\|^2 = \hat{\boldsymbol{\theta}}_t^T \boldsymbol{P_\theta}^T(\boldsymbol{I} + \eta^2\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{P_\theta}\hat{\boldsymbol{\theta}}_t + \hat{\boldsymbol{\phi}}_t^T \boldsymbol{P_\phi}^T(\boldsymbol{I} + \eta^2\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{P_\phi}\hat{\boldsymbol{\phi}}_t \tag{47}$$

635 Now using the fact that $\boldsymbol{P_\theta} = \boldsymbol{P_\theta}^T$, $\boldsymbol{P_\phi} = \boldsymbol{P_\phi}^T$, we can write

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2 \leq \frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}\boldsymbol{B}^T)} + \frac{\rho^2(1 + \eta\boldsymbol{C})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})}.$$

636 For any η that satisfying

$$\frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}\boldsymbol{B}^T)} + \frac{\rho^2(1 + \eta\boldsymbol{C})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2\lambda_{\min}(\boldsymbol{B}^T\boldsymbol{B})} < \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi_t} - \boldsymbol{\phi^*}\|^2 \tag{48}$$

637 we have that

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi_{t+1}} - \boldsymbol{\phi^*}\|^2 < \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi_t} - \boldsymbol{\phi^*}\|^2 \tag{49}$$

638 i.e., SPPM converges asymptotically towards $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$. $\qquad\square$

A.5 Proof of Theorem 5.2

640 *Proof.* In order to proof the convergence of SPPM in bilinear games, we first show that the SPPM
641 update rule is equivalent to that of the following Proximal Point Method:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_{t+1}) \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta\nabla_\phi f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_{t+1}) \end{cases} \tag{50}$$

642 In the Bilinear game, the SPPM update is:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_{t+1}) \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta\nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_{t+1}) \end{cases}$$
$$= \begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{M}\boldsymbol{\phi}_{t+1} \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta\boldsymbol{M}^T\boldsymbol{\theta}_{t+1} \end{cases}$$

23

and the PPM update is:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_\theta f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_{t+1}) \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \nabla_\phi f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_{t+1}) \end{cases}$$

$$= \begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{M} \boldsymbol{\phi}_{t+1} \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \boldsymbol{M}^T \boldsymbol{\theta}_{t+1} \end{cases}$$

Thus SPPM and PPM are equivalent in the Bilinear game. The convergence result of PPM in bilinear games has been proved in Theorem 2 of [49]:

Theorem A.3. *Consider the Bilinear game and the PPM method. Further, we define $r_t = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2$. Then, for any $\eta > 0$, the iterates $\{\boldsymbol{\theta}_t, \boldsymbol{\phi}_t\}_{t \geq 0}$ generated by SPPM satisfy*

$$r_{t+1} \leq \frac{1}{1 + \eta^2 \lambda_{\min}(\boldsymbol{M}^T \boldsymbol{M})} r_t. \tag{51}$$

Therefore, SPPM and PPM have the same convergence property in bilinear games. $\qquad\square$

A.6 Proof of Theorem 5.3

Proof. Consider the learning dynamics:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_{t+1})$$
$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \nabla_\theta f(\boldsymbol{\theta}_{t+1}, \boldsymbol{\phi}_t)$$

In the Quadratic game, the SPPM update rule can be written as:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{A} \boldsymbol{\theta}_t - \boldsymbol{C} \boldsymbol{\phi}_{t+1} \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \boldsymbol{B} \boldsymbol{\phi}_t + \boldsymbol{C}^T \boldsymbol{\theta}_{t+1} \end{cases} \tag{52}$$

Then we can rewrite the learning dynamics:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{A} \boldsymbol{\theta}_t - \eta \boldsymbol{C} \boldsymbol{\phi}_{t+1}$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{A} \boldsymbol{\theta}_t - \eta \boldsymbol{C}(\boldsymbol{\phi}_t + \eta \boldsymbol{C}^T \boldsymbol{\theta}_{t+1} + \eta \boldsymbol{B} \boldsymbol{\phi}_t)$$
$$(\boldsymbol{I} + \eta^2 \boldsymbol{C}\boldsymbol{C}^T)\boldsymbol{\theta}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{A})\boldsymbol{\theta}_t - \eta \boldsymbol{C}(\boldsymbol{I} + \eta \boldsymbol{B})\boldsymbol{\phi}_t$$
$$\boldsymbol{\theta}_{t+1} = (\boldsymbol{I} + \eta^2 \boldsymbol{C}\boldsymbol{C}^T)^{-1}\left[(\boldsymbol{I} - \eta \boldsymbol{A})\boldsymbol{\theta}_t - \eta \boldsymbol{C}(\boldsymbol{I} + \eta \boldsymbol{B})\boldsymbol{\phi}_t\right] \tag{53}$$

Similarly, for the other player we have

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \boldsymbol{C}^T \boldsymbol{\theta}_{t+1} + \eta \boldsymbol{B} \boldsymbol{\phi}_t$$
$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta \boldsymbol{C}^T(\boldsymbol{\theta}_t - \eta \boldsymbol{A} \boldsymbol{\theta}_t - \eta \boldsymbol{C} \boldsymbol{\phi}_{t+1}) + \eta \boldsymbol{B} \boldsymbol{\phi}_t$$
$$(\boldsymbol{I} + \eta^2 \boldsymbol{C}^T \boldsymbol{C})\boldsymbol{\phi}_{t+1} = \eta \boldsymbol{C}^T(\boldsymbol{I} - \eta \boldsymbol{A})\boldsymbol{\theta}_t + (\boldsymbol{I} + \eta \boldsymbol{B})\boldsymbol{\phi}_t$$
$$\boldsymbol{\phi}_{t+1} = (\boldsymbol{I} + \eta^2 \boldsymbol{C}^T \boldsymbol{C})^{-1}\left[\eta \boldsymbol{C}^T(\boldsymbol{I} - \eta \boldsymbol{A})\boldsymbol{\theta}_t + (\boldsymbol{I} + \eta \boldsymbol{B})\boldsymbol{\phi}_t\right] \tag{54}$$

Let us define the symmetric matrices $\boldsymbol{Q_\theta} = (\boldsymbol{I} + \eta^2 \boldsymbol{C}\boldsymbol{C}^T)^{-1}$, $\boldsymbol{Q_\phi} = (\boldsymbol{I} + \eta^2 \boldsymbol{C}^T \boldsymbol{C})^{-1}$ and $\boldsymbol{P_\theta} = (\boldsymbol{I} - \eta \boldsymbol{A})$, $\boldsymbol{P_\phi} = (\boldsymbol{I} + \eta \boldsymbol{B})$. Further we define $r_t = \|\boldsymbol{\theta}_{t+1}\|^2 + \|\boldsymbol{\phi}_{t+1}\|^2$. Based on these definitions, and the expressions in (53) and (54) we have

$$\|\boldsymbol{\theta}_{t+1}\|^2 + \|\boldsymbol{\phi}_{t+1}\|^2 = \|\boldsymbol{Q_\theta} \boldsymbol{P_\theta} \boldsymbol{\theta}_t\|^2 + \eta^2 \|\boldsymbol{Q_\theta} \boldsymbol{C} \boldsymbol{P_\phi} \boldsymbol{\phi}_t\|^2 + \|\boldsymbol{Q_\phi} \boldsymbol{C}^T \boldsymbol{P_\theta} \boldsymbol{\theta}_t\|^2 + \|\boldsymbol{Q_\phi} \boldsymbol{P_\phi} \boldsymbol{\phi}_t\|^2$$
$$- 2\eta \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T \boldsymbol{Q_\theta}^T \boldsymbol{Q_\theta} \boldsymbol{C} \boldsymbol{P_\phi} \boldsymbol{\phi}_t + 2\eta \boldsymbol{\phi}_t^T \boldsymbol{P_\phi}^T \boldsymbol{Q_\phi}^T \boldsymbol{Q_\phi} \boldsymbol{C}^T \boldsymbol{P_\theta} \boldsymbol{\theta}_t \tag{55}$$

To simplify the expression in (55) we use Lemma A.1 to obtain the following equations:

$$\boldsymbol{Q_\theta} \boldsymbol{C} = \boldsymbol{C} \boldsymbol{Q_\phi} \tag{56}$$
$$\boldsymbol{Q_\phi} \boldsymbol{C}^T = \boldsymbol{C}^T \boldsymbol{Q_\theta} \tag{57}$$

Using this lemma, we can show that

$$\boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T \boldsymbol{Q_\theta}^T \boldsymbol{Q_\theta} \boldsymbol{C} \boldsymbol{P_\phi} \boldsymbol{\phi}_t = \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T \boldsymbol{Q_\theta}^T \boldsymbol{C} \boldsymbol{Q_\phi} \boldsymbol{P_\phi} \boldsymbol{\phi}_t = \boldsymbol{\phi}_t^T \boldsymbol{P_\phi}^T \boldsymbol{Q_\phi}^T \boldsymbol{C}^T \boldsymbol{Q_\theta} \boldsymbol{P_\theta} \boldsymbol{\theta}_t = \boldsymbol{\phi}_t^T \boldsymbol{P_\phi}^T \boldsymbol{Q_\phi}^T \boldsymbol{Q_\phi} \boldsymbol{C}^T \boldsymbol{P_\theta} \boldsymbol{\theta}_t$$

24

where the intermediate equality holds as $\boldsymbol{a}^T \boldsymbol{C} = \boldsymbol{C}^T \boldsymbol{a}$. Hence, the expression in (55) can be simplified as

$$\|\boldsymbol{\theta}_{t+1}\|^2 + \|\boldsymbol{\phi}_{t+1}\|^2 = \|\boldsymbol{Q_\theta P_\theta \theta}_t\|^2 + \eta^2 \|\boldsymbol{Q_\theta C P_\phi \phi}_t\|^2 + \|\boldsymbol{Q_\phi C}^T \boldsymbol{P_\theta \theta}_t\|^2 + \|\boldsymbol{Q_\phi P_\phi \phi}_t\|^2 \quad (58)$$

We simplify equation (58) as follows. Consider the term involving $\boldsymbol{\theta}_t$. We have

$$\begin{aligned}
\|\boldsymbol{Q_\theta P_\theta \theta}_t\|^2 + \eta^2 \|\boldsymbol{Q_\phi C}^T \boldsymbol{P_\theta \theta}_t\|^2 &= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T \boldsymbol{Q_\theta^2 P_\theta \theta}_t + \eta^2 \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T \boldsymbol{C Q_\phi^2 C}^T \boldsymbol{P_\theta \theta}_t \\
&= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{Q_\theta^2} + \eta^2 \boldsymbol{C Q_\phi^2 C}^T) \boldsymbol{P_\theta \theta}_t \\
&= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{Q_\theta^2} + \eta^2 \boldsymbol{C Q_\phi C}^T \boldsymbol{Q_\theta}) \boldsymbol{P_\theta \theta}_t \\
&= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{Q_\theta^2} + \eta^2 \boldsymbol{C C}^T \boldsymbol{Q_\theta Q_\theta}) \boldsymbol{P_\theta \theta}_t \\
&= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{I} + \eta^2 \boldsymbol{C C}^T) \boldsymbol{Q_\theta^2 P_\theta \theta}_t \\
&= \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{I} + \eta^2 \boldsymbol{C C}^T)^{-1} \boldsymbol{P_\theta \theta}_t \quad (59)
\end{aligned}$$

where the last equality follows by replacing $\boldsymbol{Q_\theta}$ by its definition. The same procedure follows for the term involving $\boldsymbol{\phi}_t$ which leads to the expression

$$\|\boldsymbol{Q_\phi P_\phi \phi}_t\|^2 + \eta^2 \|\boldsymbol{Q_\theta C P_\phi \phi}_t\|^2 = \boldsymbol{\phi}_t^T \boldsymbol{P_\phi}^T (\boldsymbol{I} + \eta^2 \boldsymbol{C}^T \boldsymbol{C})^{-1} \boldsymbol{P_\phi \phi}_t. \quad (60)$$

Substitute $\|\boldsymbol{Q_\theta P_\theta \theta}_t\|^2 + \eta^2 \|\boldsymbol{Q_\phi C}^T \boldsymbol{P_\theta \theta}_t\|^2$ and $\|\boldsymbol{Q_\phi P_\phi \phi}_t\|^2 + \eta^2 \|\boldsymbol{Q_\theta C P_\phi \phi}_t\|^2$ in (58) with the expressions in (59) and (60), respectively, to obtain

$$\|\boldsymbol{\theta}_{t+1}\|^2 + \|\boldsymbol{\phi}_{t+1}\|^2 = \boldsymbol{\theta}_t^T \boldsymbol{P_\theta}^T (\boldsymbol{I} + \eta^2 \boldsymbol{C C}^T)^{-1} \boldsymbol{P_\theta \theta}_t + \boldsymbol{\phi}_t^T \boldsymbol{P_\phi}^T (\boldsymbol{I} + \eta^2 \boldsymbol{C}^T \boldsymbol{C})^{-1} \boldsymbol{P_\phi \phi}_t. \quad (61)$$

Now using the expression in (61) and the fact that $\boldsymbol{P_\theta} = \boldsymbol{P_\theta}^T$, $\boldsymbol{P_\phi} = \boldsymbol{P_\phi}^T$ and $\lambda_{\min}(\boldsymbol{C}^T \boldsymbol{C}) = \lambda_{\min}(\boldsymbol{C C}^T)$, we can write

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}^*\|^2 \leq \frac{\rho^2(\boldsymbol{I} - \eta\boldsymbol{A})\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + \rho^2(1 + \eta\boldsymbol{B})\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2}{\boldsymbol{I} + \eta^2 \lambda_{\min}(\boldsymbol{C}^T \boldsymbol{C})}.$$

$\square$

A.7 Competitive Gradient Descent as an Approximation of SPPM

In this section, we justify our results in Section 4.1 that Competitive Gradient Descent is a first order Taylor approximation of SPPM. Firstly, we consider the standard definition of CGD:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\boldsymbol{I} + \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))^{-1} (\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \eta\nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)) \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta(\boldsymbol{I} + \eta^2 \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))^{-1} (\nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta\nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)) \end{cases}$$

Rewriting the update rules we can get:

$$(\boldsymbol{I} + \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) = -\eta(\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \eta\nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$$

Similarly, we have:

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}_t)$$

Therefore, CGD is a first order approximation of SPPM. Then we prove that the standard definition of CGD is equivalent to the update rule in Table 1. Consider the update rule in Table 1 and its footnote, we have:

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta\nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}_t) \\ \boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta\nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \eta\nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \end{cases} \quad (62)$$

Substituting $(\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}_t)$ into the first equation of (62) we get:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta\nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\eta\nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \eta\nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t))$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$$

$$(\boldsymbol{I} + \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) = -\eta\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\boldsymbol{I} + \eta^2 \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))^{-1} (\nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \eta\nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t)).$$

Substituting $(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$ into the second equation of (62) and applying similar arguments we get:

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t + \eta(\boldsymbol{I} + \eta^2 \nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_{\theta\phi} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))^{-1} (\nabla_\phi f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) - \eta\nabla_{\phi\theta} f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t) \nabla_\theta f(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t))$$

Thus the update rule in Table 1 is equivalent to the standard definition of CGD and it is equivalent to the first order Taylor approximation of SPPM.

25

681 ## A.8 Experiments on Bilinear and Quadratic Games

682 **Bilinear Game** Consider the following bilinear game:

$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} a\theta\phi \tag{63}$$

683 the example presented in Figure 1 is an example of using different algorithms to solve the bilinear
684 game above with coefficient $a = 10$. For sake of completeness, we also provide a grid of experiment
685 results for different algorithms with different coefficients a and learning rates η, starting from the
same point $(\theta_0, \phi_0) = (-12, 10)$. The result is presented in Figure 6. The experiment demonstrates
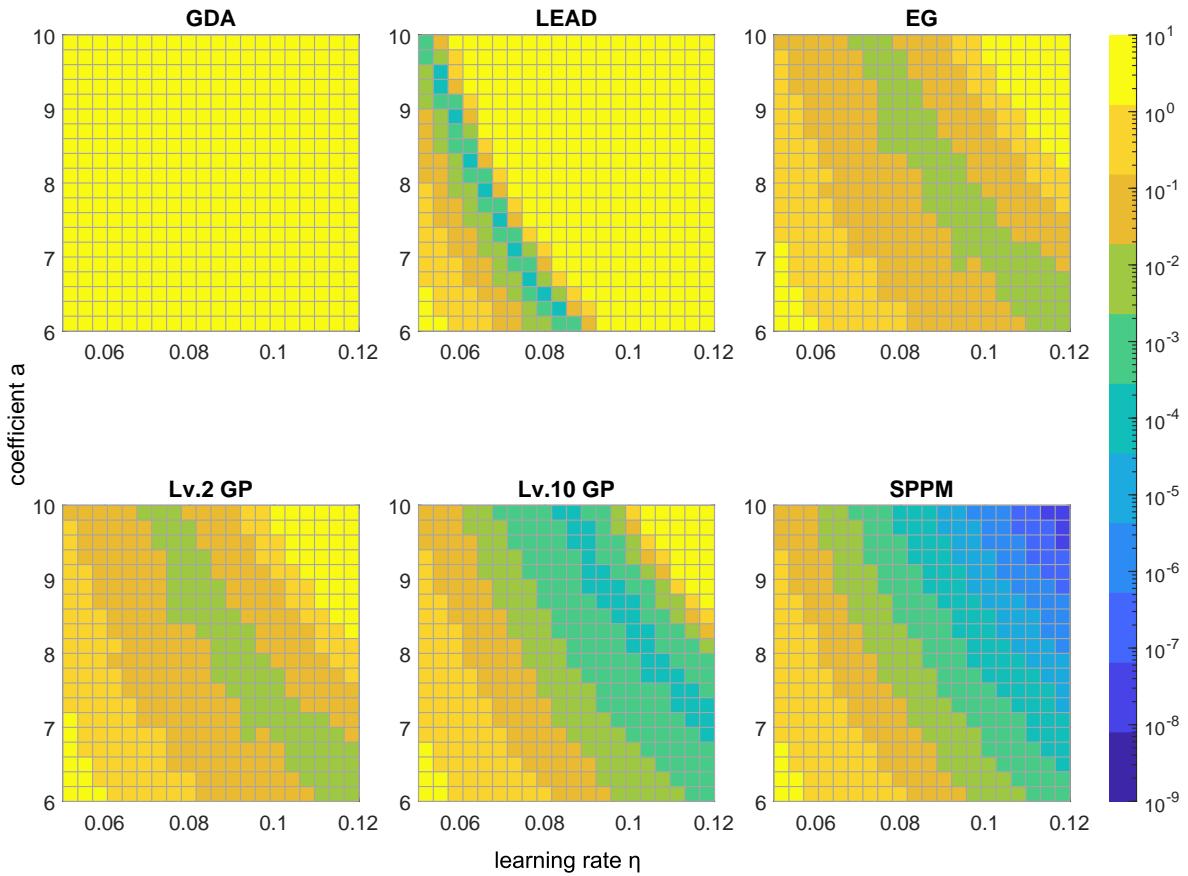
Figure 6: A grid of experiments on the bilinear game for different algorithms with different values of coefficient a and learning rates η. The color in each cell indicates the distance to the equilibrium after 50 iterations.

686
687 that, in a bilinear game, Lv.2 GP is equivalent to the extra-gradient method, and higher level Lv.k GP
688 performs better with increased coefficient a and learning rate η as long as it remains a contraction
689 (i.e., $\eta < a^{-1}$).

26

690 **Quadratic Game** For the quadratic game presented in Figure 3, we randomly initialize the matrices
691 $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A} = \begin{bmatrix} 1.8398 & 0.5195 & 1.2537 & 1.7470 & 1.2769 \\ 0.5195 & 0.6586 & 0.4476 & 0.8898 & 1.1309 \\ 1.2537 & 0.4476 & 1.4440 & 1.3923 & 0.8877 \\ 1.7470 & 0.8898 & 1.3923 & 2.1249 & 1.7664 \\ 1.2769 & 1.1309 & 0.8877 & 1.7664 & 2.1553 \end{bmatrix} \quad \boldsymbol{B} = - \begin{bmatrix} 1.0821 & 1.2427 & 1.0093 & 1.3335 & 0.6761 \\ 1.2427 & 2.2031 & 1.3236 & 1.8566 & 0.9394 \\ 1.0093 & 1.3236 & 1.2393 & 1.3675 & 0.9065 \\ 1.3335 & 1.8566 & 1.3675 & 1.9081 & 0.9693 \\ 0.6761 & 0.9394 & 0.9065 & 0.9693 & 0.7141 \end{bmatrix}$$

692 where $\boldsymbol{A}$ is symmetric and positive definite and $\boldsymbol{B}$ is symmetric and negative definite. The interaction
693 matrix is defined as:

$$\boldsymbol{C} = \begin{bmatrix} c & 0 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & 0 \\ 0 & 0 & c & 0 & 0 \\ 0 & 0 & 0 & c & 0 \\ 0 & 0 & 0 & 0 & c \end{bmatrix}$$

694 where c represents the strength of the interaction between the two players. The starting point $\boldsymbol{\theta}_0$ and
695 $\boldsymbol{\phi}_0$ are $[0.1270, 0.9667, 0.2605, 0.8972, 0.3767]^T$ and $[0.3362, 0.4514, 0.8403, 0.1231, 0.5430]^T$ respectively.
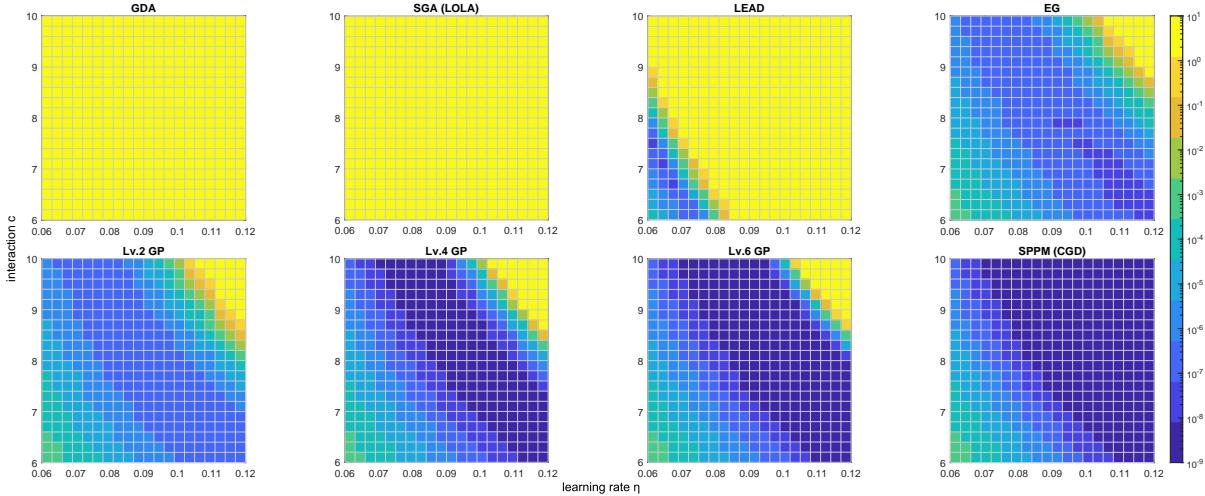
Figure 7: A grid of experiments on the quadratic game for different algorithms with different values of coefficient a and learning rates η. The color in each cell indicates the distance to the equilibrium after 50 iterations.

696

A.9 Experiments on 8-Gaussians

698 **Dataset** The target distribution is a mixture of 8-Gaussians with standard deviation equal to 0.05
699 and modes uniformly distributed around a unit circle.

700 **Architecture** The architecture for the generator and the discriminator, each consists of four hidden
701 layers followed by ReLU activation function. The weight initialization uses PyTorch's default
702 initialization scheme. The architecture is presented in Table 8:

703 **Experiment** For our experiments, we used the PyTorch framework. Furthermore, the batch size we
704 used is 128.

27

Table 5: Architecture used for the Mixture of 8-Gaussians

Generator	Discriminator
Input: $z \in \mathbb{R}^{64} \sim \mathcal{N}(0, I)$	*Input: $x \in \mathbb{R}^2$*
Linear($64 \to 2000$)	Linear($2 \to 2000$)
ReLU	ReLU
Linear($2000 \to 2000$)	Linear($2000 \to 2000$)
ReLU	ReLU
Linear($2000 \to 2000$)	Linear($2000 \to 2000$)
ReLU	ReLU
Linear($2000 \to 1$)	Linear($2000 \to 1$)

A.10 Experiments on CIFAR-10 and STL-10

For our experiments, we used the PyTorch[2] framework. For experiments on CIFAR-10 and STL-10, we compute the FID and IS metrics using the provided implementations in Tensorflow[3] for consistency with related works.
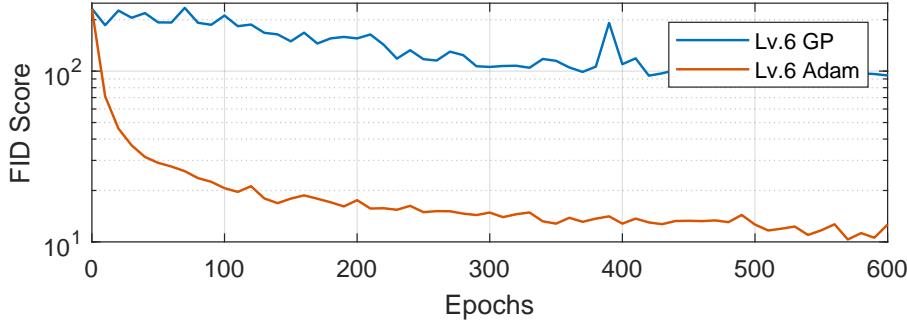
Figure 8: Comparison between Lv.k GP and Lv.k Adam on generating CIFAR-10 images. We can see significant improvements in FID when using the Lv.k Adam algorithm we proposed.

Lv.k GP vs Lv.k Adam In experiments, we compare the performance of Lv.k GP and Lv.k Adam on the task of CIFAR-10 image generation. The experiment results is presented in Figure 8. The experiments on Lv.k GP and Lv.k Adam use the same initialization and hyperparameters. According to our experiments, Lv.k Adam converges much faster than Lv.k GP for the same choice of k and learning rates.

Adam vs Lv.k Adam We also present a comparison between the performance of Adam and Lv.k Adam optimizers on the task of STL-10 image generation. The experiment results is presented in Figure 9. Under the same choice of hyperparameters and identical model parameter initialization, Lv.k Adam consistently outperforms the Adam optimizer in terms of FID score.

Accelerated Lv.k Adam In this section, we propose an accelerated version of Lv.k Adam. The intuition is that we update the min player θ and the max player ϕ in an alternating order. The corresponding Lv.k GP algorithm can be writen as:

$$\text{Reasoning: } \begin{cases} \theta_t^{(k)} = \theta_t - \eta \nabla_{\theta} f(\theta_t, \phi_t^{(k-1)}) \\ \phi_t^{(k)} = \phi_t - \eta \nabla_{\phi} g(\theta_t^{(k)}, \phi_t) \end{cases} \text{Update: } \begin{cases} \theta_{t+1} = \theta_t - \eta \nabla_{\theta} f(\theta_t, \phi_t^{(k)}) \\ \phi_{t+1} = \phi_t - \eta \nabla_{\phi} g(\theta_t^{(k)}, \phi_t) \end{cases} \text{(Alt-Lv.}k\text{ GP)}$$

---

[2]https://pytorch.org/
[3]https://tensorflow.org/

28
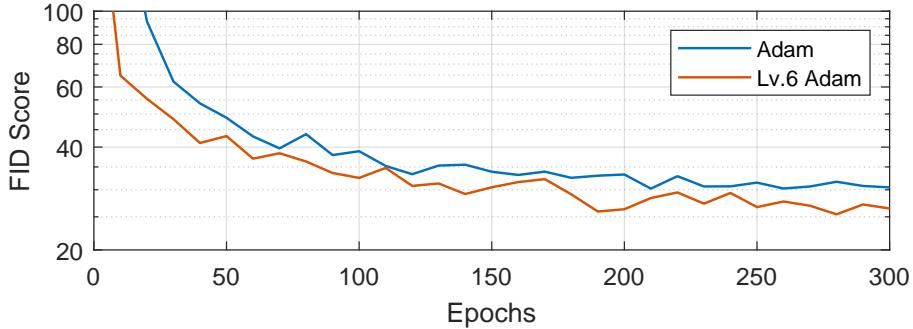
Figure 9: Comparison between Adam and Lv.k Adam on generating STL-10 images. We can see that Lv.k Adam consistently outperforms the Adam optimizer in terms of FID score.

Instead of responding to $\boldsymbol{\theta}_t^{(k-1)}$, in Alt-Lv.$k$ GP, the max player $\boldsymbol{\phi}_t^{(k)}$ acts in response to the min player's current action, $\boldsymbol{\theta}_t^{(k)}$. A Lv.$k$ min player in Alt-Lv.k GP is equivalent to a Lv.$2k-1$ player in the Lv.k GP, and a Lv.k max player in Alt-Lv.k GP is equivalent to a Lv.$2k$ player in the Lv.k GP, respectively. Therefore, it is easy to verify that Alt-Lv.k GP converges two times faster than Lv.k GP and the corresponding Alt-Lv.k Adam algorithm is provided in Algorithm 2.

Architecture In this section, we describe the model we used to evaluate the performance of Lv.k Adam for generating CIFAR-10[4] and STL-10 datasets. With 'conv' we denote a convolutional layer and 'transposed conv' a transposed convolution layer. The models use Batch Normalization and Spectral Normalization. The model's parameters are initialized with Xavier initialization.

Images generated on CIFAR-10 and STL-10 In this section, we present sample images generated by the best performing trained generators on CIFAR-10 and STL-10.

---

[4]CIFAR10 is released under the MIT license.

29

Algorithm 2: Accelerated Level k Adam: proposed Adam with recursive reasoning steps

---

Input: Stopping time T, reasoning steps k, learning rate $\eta_{\boldsymbol{\theta}}, \eta_{\boldsymbol{\phi}}$, decay rates for momentum
estimates β_1, β_2, initial weight $(\boldsymbol{\theta}_0, \boldsymbol{\phi}_0)$, $\boldsymbol{P_x}$ and $\boldsymbol{P_z}$ real and noise-data distributions,
losses $\mathcal{L}_G(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{x}, \boldsymbol{z})$ and $\mathcal{L}_D(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{x}, \boldsymbol{z})$, $\epsilon = 1e - 8$.

Parameters: Initial parameters: $\boldsymbol{\theta}_0, \boldsymbol{\phi}_0$
                 Initialize first moments: $\boldsymbol{m}_{\theta,0} \leftarrow 0, \boldsymbol{m}_{\phi,0} \leftarrow 0$
                 Initialize second moments: $\boldsymbol{v}_{\theta,0} \leftarrow 0, \boldsymbol{v}_{\phi,0} \leftarrow 0$

for $t=0,\ldots,T\text{-}1$ **do**
     Sample new mini-batch: $\boldsymbol{x}, \boldsymbol{z} \sim \boldsymbol{P_x}, \boldsymbol{P_z}$,
     $\boldsymbol{\theta}_t^{(0)} \leftarrow \boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(0)} \leftarrow \boldsymbol{\phi}_t$,
     for $n=1,\ldots,k$ **do**
         Compute stochastic gradient: $\boldsymbol{g}_{\boldsymbol{\theta},t}^{(n)} = \nabla_{\theta} \mathcal{L}_G(\boldsymbol{\theta}_t, \boldsymbol{\phi}_t^{(n-1)}, \boldsymbol{x}, \boldsymbol{z})$;

         Update estimate of first moment: $\boldsymbol{m}_{\theta,t}^{(n)} = \beta_1 \boldsymbol{m}_{\theta,t-1} + (1 - \beta_1) \boldsymbol{g}_{\theta,t}^{(n)}$;

         Update estimate of second moment: $\boldsymbol{v}_{\theta,t}^{(n)} = \beta_2 \boldsymbol{v}_{\theta,t-1} + (1 - \beta_2)(\boldsymbol{g}_{\theta,t}^{(n)})^2$;

         Correct the bias for the moments: $\hat{\boldsymbol{m}}_{\theta,t}^{(n)} = \frac{\boldsymbol{m}_{\theta,t}^{(n)}}{(1-\beta_1^t)}, \hat{\boldsymbol{v}}_{\theta,t}^{(n)} = \frac{\boldsymbol{v}_{\theta,t}^{(n)}}{(1-\beta_2^t)}$;

         Perform Adam update: $\boldsymbol{\theta}_t^{(n)} = \boldsymbol{\theta}_t - \eta_\theta \frac{\hat{\boldsymbol{m}}_{\theta,t}^{(n)}}{\sqrt{\hat{\boldsymbol{v}}_{\theta,t}^{(n)}} + \epsilon}$;

         Compute stochastic gradient: $\boldsymbol{g}_{\boldsymbol{\phi},t}^{(n)} = \nabla_{\phi} \mathcal{L}_D(\boldsymbol{\theta}_t^{(n)}, \boldsymbol{\phi}_t, \boldsymbol{x}, \boldsymbol{z})$;

         Update estimate of first moment: $\boldsymbol{m}_{\phi,t}^{(n)} = \beta_1 \boldsymbol{m}_{\phi,t-1} + (1 - \beta_1) \boldsymbol{g}_{\phi,t}^{(n)}$;

         Update estimate of second moment: $\boldsymbol{v}_{\phi,t}^{(n)} = \beta_2 \boldsymbol{v}_{\phi,t-1} + (1 - \beta_2)(\boldsymbol{g}_{\phi,t}^{(n)})^2$;

         Correct the bias for the moments: $\hat{\boldsymbol{m}}_{\phi,t}^{(n)} = \frac{\boldsymbol{m}_{\phi,t}^{(n)}}{(1-\beta_1^t)}, \hat{\boldsymbol{v}}_{\phi,t}^{(n)} = \frac{\boldsymbol{v}_{\phi,t}^{(n)}}{(1-\beta_2^t)}$;

         Perform Adam update: $\boldsymbol{\phi}_t^{(n)} = \boldsymbol{\phi}_t - \eta_\phi \frac{\hat{\boldsymbol{m}}_{\phi,t}^{(n)}}{\sqrt{\hat{\boldsymbol{v}}_{\phi,t}^{(n)}} + \epsilon}$;

     $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t^{(k)}, \boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t^{(k)}$;
     $\boldsymbol{m}_{\theta,t} \leftarrow \boldsymbol{m}_{\theta,t}^{(k)}, \boldsymbol{m}_{\phi,t} \leftarrow \boldsymbol{m}_{\phi,t}^{(k)}$;
     $\boldsymbol{v}_{\theta,t} \leftarrow \boldsymbol{v}_{\theta,t}^{(k)}, \boldsymbol{v}_{\phi,t} \leftarrow \boldsymbol{v}_{\phi,t}^{(k)}$

Table 6: ResNet blocks used for the SN-GAN architectures on CIFAR-10 image generation, for the generator (left) and the discriminator (right).

G-ResBlock	D-ResBlock ($l-$th block)
Shortcut:	*Shortcut:*
Upsample($\times 2$)	[AvgPool (ker: 2×2)], if $l = 1$
Residual:	conv (ker: 1×1, $3_{l=1}/128_{l\neq1} \to 128$; stride: 1)
Batch Normalization	Spectral Normalization
ReLU	[AvgPool (ker: 2×2, stride: 2)], if $l = 1$
Upsample($\times 2$)	*Residual:*
conv (ker:3×3, $256 \to 256$; stride: 1; pad: 1)	[ReLU], if $l \neq 1$
Batch Normalization	conv (ker: 3×3, $3_{l=1}/128_{l\neq1} \to 128$; stride: 1; pad: 1)
ReLU	Spectral Normalization
conv (ker:3×3, $256 \to 256$; stride: 1; pad: 1)	ReLU
	conv (ker: 1×1, $128 \to 128$; stride: 1)
	Spectral Normalization
	AvgPool (ker: 2×2)

30

Table 7: SN-GAN architectures for experiments on CIFAR-10

Generator	Discriminator
Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$	*Input:* $x \in \mathbb{R}^{3 \times 32 \times 32}$
Linear($128 \rightarrow 4096$)	D-ResBlock
G-ResBlock	D-ResBlock
G-ResBlock	D-ResBlock
G-ResBlock	D-ResBlock
Batch Normalization	ReLU
ReLU	AvgPool(ker:8×8)
conv (ker:3×3, $256 \rightarrow 3$; stride: 1; pad: 1)	Linear($128 \rightarrow 1$)
Tanh	Spectral Normalization

Table 8: SN-GAN architectures for experiments on STL-10

Generator	Discriminator
Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$	*Input:* $x \in \mathbb{R}^{3 \times 48 \times 48}$
Linear($128 \rightarrow 6 \times 6 \times 512$)	D-ResBlock down $64 \rightarrow 128$
G-ResBlock up $512 \rightarrow 256$	D-ResBlock down $3 \rightarrow 128$
G-ResBlock up $256 \rightarrow 128$	D-ResBlock down $128 \rightarrow 256$
G-ResBlock up $128 \rightarrow 64$	D-ResBlock down $256 \rightarrow 512$
Batch Normalization	D-ResBlock $512 \rightarrow 1024$
ReLU	ReLU, AvgPool (ker: 8×8)
conv (ker: 3×3, $64 \rightarrow 3$; stride: 1; pad: 1)	Linear($128 \rightarrow 1$)
Tanh	Spectral Normalization

31

Figure 10: The presented samples are generated by the best performing trained generator on CIFAR-10, using Lv.6 Adam. This gives a FID score of 10.12.

32
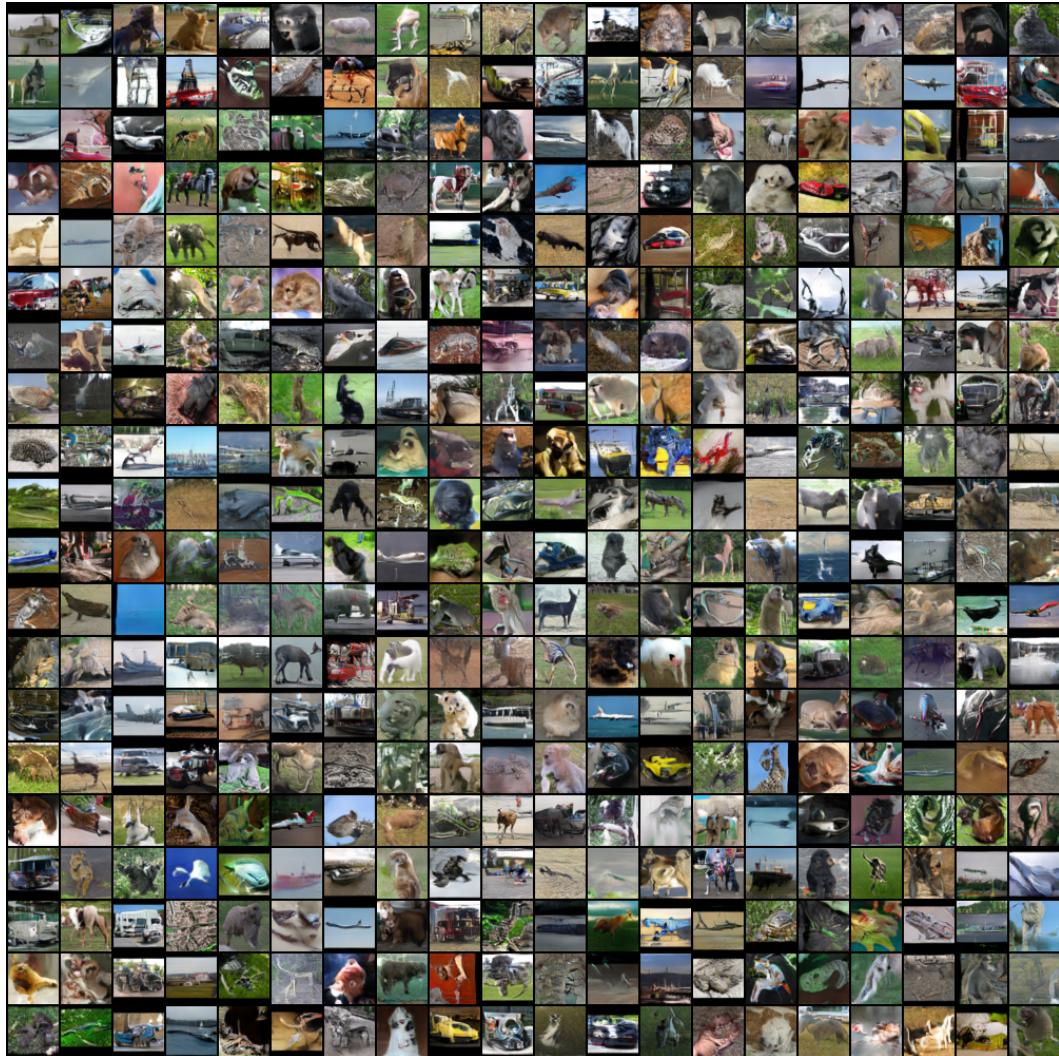
Figure 11: The presented samples are generated by the best performing trained generator on STL-10, using Lv.6 Adam. This gives a FID score of 25.43.

33